Washington Assessment of Student Learning

Grade 10

2000

Technical Report

Prepared by
Catherine S. Taylor and Yoonsun Lee
University of Washington

for
Office of the Superintendent of Public Instruction
P.O. Box 47220
Olympia, Washington 98504-7220

October 18, 2001

TABLE OF CONTENTS

TABLE OF CONTENTS (Cont.)

TABLE OF TABLES

TABLE OF TABLES (Cont.)

TABLE OF TABLES (Cont.)

TABLE OF TABLES (Cont.)

TABLE OF FIGURES

The Washington Assessment of Student Learning: March 2000

PURPOSE OF TECHNICAL REPORT

Standards for Educational and Psychological Testing (AERA/APA/NCME, 1999) require that test developers and publishers produce a technical manual. The technical manual must provide overall information documenting the technical quality of the assessment, including evidence for the reliability and validity of test scores. This document contains the technical information for the 2000 *Washington Assessment of Student Learning*: Grade 10 Assessment for Reading, Mathematics, Listening and Writing.

PART 1

OVERVIEW

BACKGROUND FOR THE STATE ASSESSMENT PROGRAM

In 1993, Washington State embarked on the development of a comprehensive school change effort that has as its primary goal the improvement of teaching and learning. Created by the state legislature in 1993, the Commission on Student Learning was charged with three important tasks in support of this school change effort:

- to establish Essential Academic Learning Requirements (EALRs) that describe what all students should know and be able to do in eight content areas--reading, writing, communication, mathematics, science, health/fitness, social studies, and the arts;
- to develop an assessment system to measure student progress at three grade levels towards achieving the EALRs; and
- to recommend an accountability system that recognizes and rewards successful schools and provides support and assistance to less successful schools.

The Commission has achieved its first major task. The EALRs in Reading, Writing, Communications, and Mathematics were first adopted in 1995 and revised in 1997 (See http://www.k12.wa.us/curriculuminstruct/EALRs.asp for the EALRs in all subject areas). Performance "benchmarks" were also established at three grade levels--elementary (Grade 4), middle (Grade 7), and high school (Grade 10). The EALRs for Science, Social Studies, Health/Fitness, and the Arts were initially adopted in 1996 and also revised in 1997. Performance "benchmarks" for these subject areas were also established at three levels – elementary, middle, and high school.

The Commission's second major task was to develop an assessment system to determine the extent to which students are achieving the knowledge and skills defined by the EALRs. The assessments for Reading, Writing, Communication, and Mathematics have been developed at Grades 4 and 7 and were both operational as of spring, 1998. The Grade 10 assessment in these same content areas was pilot-tested in spring, 1998 and was operational beginning spring, 1999. Participation in the Grade 4 assessment was mandatory for all public schools beginning spring, 1998. Participation in the Grade 7 and 10 assessments was voluntary until spring, 2001.

Work is underway to develop middle and high school assessments in Science beginning with pilot assessments in spring, 1998 and operational assessments in spring, 2003. Grade 5 science assessments will be piloted in 2003 and operational in 2004. Assessment development work in the other content areas – Social Studies, Health and Fitness, and the Arts – are expected to be operational in 2005 or 2006.

WASHINGTON ASSESSMENT SYSTEM

The assessment system has four major components: state-level assessments, classroom-based assessments, professional staff development, and school and system context indicators. These components are described briefly below. Two additional features, the Certificate of Mastery and the Accountability System, are also briefly described.

State-Level Assessments in Reading, Writing, Listening, and Mathematics

The state-level assessments require students to both select and create answers to demonstrate their knowledge, skills, and understanding in each of the EALRs – from multiple-choice and short-answer items to more extended responses, essays, and problem solving tasks. Student, school, and district scores are reported for the operational assessments. The state-level operational test forms are standardized and "on demand", meaning that all students respond to the same items, under the same conditions, and at the same time during the school year.

All of the state-level assessments are untimed; that is, students may have as much time as they reasonably need to complete their work. Guidelines for providing accommodations to students with special needs have been developed to encourage the inclusion of as many students as possible. Special needs students include those in special education programs, those with Section 504 plans, English language learners (ESL/bilingual), migrant students, and highly capable students. A broad range of accommodations allows nearly all students access to some or all parts of the assessment (see *Guidelines for Inclusion and Accommodations for Special Populations on State-Level Assessments*).

Classroom teachers and curriculum specialists from across Washington were selected to assist with the development of the items for the state-level assessments. Two content committees were created at each grade level--one for Reading/Writing/Communication and one for Mathematics. Working with content and assessment specialists from the Riverside Publishing Company (one of the Commission's assessment development contractors), these committees defined the test and item specifications consistent with the Washington State Essential Academic Learning Requirements, reviewed all items prior to pilot testing, and provided final review and approval of all items after pilot testing. A separate "fairness" committee, composed of individuals reflective of Washington's diversity, also reviewed all items for words or content that might be offensive to students or parents, or might disadvantage some students for reasons unrelated to the skill or concept being assessed. (See Part 2 for a more detailed description of this process.)

Literally hundreds of items were developed and pilot-tested to create a "pool" of items. This will allow the creation of new forms of the assessment each year by sampling from the pool. Statistical "equating" procedures are used to maintain the same performance standard from year to year and to provide longitudinal comparisons across years even though different items are used.

The state-level assessments in Reading, Communication, and Mathematics include a mix of multiple-choice, short-answer, and extended-response items. Having a large pool of items provides the opportunity to vary the kinds of items from year to year so that a particular item format (e.g. multiple-choice, short-answer, or extended-response) is not always associated with the same Essential Academic Learning Requirements. (See Part 2 for more detail on the item types)

Following the first operational assessment at each grade level, a standard-setting committee determined the level of performance on each assessment that would be required for students to "meet the standard" on the Essential Academic Learning Requirements. In addition, "progress categories" above and below the standard were established in Reading and Mathematics to show growth over time as well as to give students and parents an indication of how far from the standard in these content areas a student's performance is. School and district performance on the assessments is reported in terms of the percentage of students meeting the standard and in each of the progress categories. (See Part 5 for a complete description of the standard setting process).

An *Example Test* and *Assessment Sampler* for each of the Grade 4, 7, and 10 operational assessments were created for teachers, students, and parents. The *Example Tests* along with the *Assessment Samplers* include samples of the test items, the scoring criteria for the items, and examples of student responses that have been scored. In addition to these materials, an interactive CD-ROM system called NCS Mentor for Washington provides teachers and students with another means to review the Essential Academic Learning Requirements and practice scoring student responses to items like those contained on the operational assessments.

Classroom-Based Assessment

There were a number of important reasons for including classroom-based assessment as part of the new assessment system. First, classroom-based assessments help students and their teachers better understand the Essential Academic Learning Requirements and to recognize the characteristics of quality work that define good performance for each content area. Second, classroom-based assessments provide assessment of some of the EALRs for which state-level assessment is not feasible (for example, oral presentations or group discussion). Third, classroom-based assessments offer teachers and students opportunities to gather evidence of student achievement in ways that best fit the needs and interests of individual students. Fourth, classroom-based assessments help teachers become more effective in gathering valid evidence of student learning related to the Essential Academic

Learning Requirements. And finally, good classroom-based assessments can be more sensitive to the developmental needs of students and provide the flexibility necessary to better accommodate the learning styles of children with special needs. In addition to the items that may be on the state-level assessments, classroom-based assessments can provide information from oral interviews and presentations, work products, experiments and projects, or exhibitions of student work collected over a week, a month, or the entire school year.

Classroom-based assessment *Tool Kits* have been developed for the early and middle years to provide teachers with examples of good assessment strategies. The *Tool Kits* include models for paper and pencil tasks, generic checklists of skills and traits, observation assessment strategies, simple rating scales, and generic protocols for oral communications and personal interviews. At the upper grades, classroom-based assessment strategies will also include models for developing and evaluating cross-discipline, performance-based tasks. In addition to the models, the *Tool Kits* also provide content frameworks to assist teachers, at all grade levels, to relate their classroom learning goals and instruction to the Essential Academic Learning Requirements.

Professional Development

A third major component of the new assessment system emphasizes the need for ongoing, comprehensive support and professional training for teachers and administrators to improve their understanding of the Essential Academic Learning Requirements, the characteristics of sound assessments, and effective instructional strategies that will help students reach the standards. The Commission on Student Learning established fifteen "Learning and Assessment Centers" across the state. Most are managed through Washington's nine Educational Service Districts with a few managed by school district consortia. These Centers provide professional development and support to assist school and district staff in:

1 linking teaching and curriculum to high academic standards based on the EALRs;
2 learning and applying the principles of good assessment practice;
3 using a variety of assessment techniques and strategies;
4 judging student work by applying explicit scoring criteria;
5 making instructional and curricular decisions based on reliable and valid assessment information; and
6 helping students and parents to understand the EALRs and how students can achieve them.

Context Indicators

Context indicators help teachers, parents, and the public understand and interpret student performance in relation to the environment in which teaching and learning occur. Examples of potentially useful indicators include information about faculty experience and training, instructional strategies employed, special programs for students, condition of facilities and equipment, availability of appropriate instructional materials and technology, relevant characteristics of students and the community, student attendance patterns, grade to grade transition successes, and high school dropout and graduation rates. The purpose for context information is not to explain away or excuse low performance. Rather, context indicators can provide important information to schools, policy-makers, and the public about the conditions that support or inhibit success in helping all students achieve the Essential Academic Learning Requirements.

Certificate of Mastery

Once the Essential Academic Learning Requirements and new standards are fully in place, graduating seniors will be required to earn a Certificate of Mastery to get a high school diploma. The Certificate will serve as evidence that students have achieved Washington's Essential Academic Learning Requirements by meeting the standards set for the Grade 10 assessments. Preliminary recommendations for implementing the Certificate have been forwarded to the legislature and include the recommendation that initial use should be based only on meeting the standards in Reading, Writing, Communication, Mathematics, and Science. The Certificate as a high school graduation requirement would begin with the graduating class of 2008. The Commission recommended that meeting the standards in the other content areas be treated as "endorsements" rather than as requirements once those assessments are developed and operational.

School and District Accountability System

The Academic Achievement and Accountability (A+) Commission has developed recommendations for a school and district accountability system that will recognize schools that are successful in helping their students achieve the standards on the WASL assessments. Recommendations also address the need for assistance to those schools and districts in which students are not achieving the standards. The task force recommendations are currently in draft form and are available for public review (see *A+ Commission Draft Decision Document*, August 12, 2000).

Summary

The Commission on Student Learning was committed to developing an instructionally relevant, performance-based assessment system that enhances instruction and student learning. The new assessments are based directly on the EALRs. Therefore, teachers and those who provide pre-service and in-service training to teachers should be thoroughly familiar with the EALRs and the assessments that measure them. Teachers and administrators at all grade levels need to be thinking and talking together about what they must do to prepare students to achieve the EALRs and to demonstrate their achievement on classroom-based and state-level assessments.

CRITERION-REFERENCED TESTING

The purpose of an achievement test is to determine how well a student has learned important concepts and skills. Test scores are used to make inferences about students' overall performance in a particular domain. In order to decide "how well" a student has done, some external frame of reference is needed. When we compare a student's performance to a desired performance, this is considered a criterion-referenced interpretation. When we compare a student's performance to the performance of other students, this is considered a norm-referenced interpretation.

Criterion-Referenced Tests are intended to provide a measure of the degree to which students have achieved a desired set of learning targets (desired conceptual understandings and skills) that have been identified as appropriate for a given grade or developmental level in school. Careful attention is given to making certain that the items on the test represent only the desired learning targets and that there are sufficient items for each learning target to make dependable statements about students' degree of achievement related to that target. When a standard is set for a criterion-referenced test, examinee scores are compared to the standard in order to draw inferences about whether students have attained the desired level of achievement. Scores on the test are used to make statements like, "this student meets the minimum mathematics requirements for this class," or "this student knows how to apply computational skills to solve a complex word problem."

Norm-Referenced Tests are intended to provide a general measure of some achievement domain. The primary purpose of norm-referenced tests is to make comparisons between students, schools and districts. Careful attention is given to creating items that vary in difficulty so that even the most gifted students may find that some of the items are challenging and even the student who has difficulty in school may respond correctly to some items. Items are included on the test that measure below-grade-level, on-grade-level, and above-grade-level concepts and skills. Items are spread broadly across the domain. While some norm-referenced tests provide objective-level information, items for each objective may represent concepts and skills that are not easily learned by most students until later years in school. Examinee scores on a norm-referenced test are compared to the performances of a norm-group (a representative group of students of similar age and grade). Norm groups may be local (other students in a district or state) or national (representative samples of students from throughout the United States). Scores on norm-referenced tests are used to make statements like, "this student is the best student in the class," or "this student knows mathematical concepts better than 75% of the students in the norm group."

To test all of the desired concepts and skills in a domain, testing time would be inordinately long. Well designed state or national achievement tests, whether norm-or criterion-referenced, always include samples from the domain of desired concepts and skills. Therefore, when state or national achievement tests are used, we generalize from a student's performance on the sample of items in the test and estimate how the student would perform in the domain as a whole. To have a broader measure of student achievement in some domain, it is necessary to use more than one assessment. District and classroom assessments are both useful and necessary to supplement information that is derived from state or national achievement tests.

It is possible, sometimes even desirable, to have both norm-referenced and criterion-referenced information about students' performance. The referencing scheme is best determined by the intended use of the test and this is generally determined by how the test is constructed. If tests are being used to make decisions about the success of instruction, the usefulness of an instructional or administrative program, or the degree to which students have attained a set of desired learning targets, then criterion-referenced tests and interpretations are most useful. If the tests are being used to select students for particular programs or to compare students, districts, and states, then norm-referenced tests and interpretations are useful. In some cases, both norm-referenced and criterion-referenced interpretations can be made from the same achievement measures. The *Washington Assessment of Student Learning* (WASL) state level assessment is a criterion-referenced test. Therefore, student performance should be interpreted in terms of how well students have achieved the Washington state Essential Academic Learning Requirements.

APPROPRIATE USE OF TEST SCORES

Once tests are administered, WASL performance is reported at the individual, school, and district levels. The information in these reports can be used, along with other assessment information, to help with school and district curriculum planning and classroom instructional decisions. For example, if students in a school are not performing well on the WASL Reading assessment, a careful look at the strand scores (Main Ideas and Details of Fiction; Analysis, Interpretation, & Synthesis of Fiction; Critical Thinking about Fiction; Main Ideas and Details of Nonfiction; Analysis, Interpretation, and Synthesis of Nonfiction; Critical Thinking about Nonfiction) can assist in planning instruction in future years. It may be that students as a whole are successful in comprehending and interpreting literature but are not very successful with informational text. Curriculum planning can center on how to improve materials and instruction related to informational text.

While school and district scores may be useful in curriculum and instructional planning, it is important to exercise extreme caution when interpreting individual reports. The items included on WASL tests are samples from a larger domain. Scores from one test given on a single occasion should never be used to make important decisions about students' placement, the type of instruction they receive, or retention in a given grade level in school. It is important to corroborate individual scores on WASL tests with classroom-based and other local evidence of student learning (e.g., scores from district testing programs). When making decisions about individuals, multiple sources of information should be used and multiple individuals who are familiar with the student's progress and achievement (including parents, teachers, school counselors, school psychologists, specialist teachers, and possibly even the students themselves) should be brought together to make such decisions collaboratively.

DESCRIPTION OF THE TESTS

The Grade 10 2000 forms of the Washington Assessment of Student Learning measure students' achievement of the Essential Academic Learning Requirements in Reading, Writing, Listening, and Mathematics. The following tables (Tables 1-1 to 1-4) indicate the EALRs measured by each of the four tests, the test "strands", and the number of items per strand in the 2000 test form.

Table 1-1: 2000 Grade 10, Number and Content of Listening Items

Test Strand*	Number of Items
Listens and observes to gain new information	3
Checks for understanding (paraphrasing, questioning, clarifying)	3
Analyzes media messages	2
Total No. of Items	8

* Listening EALR 1: The student uses listening and observation skills to gain understanding.

Table 1-2: 2000 Grade 10, Number and Content of Reading Items

Type of Reading Passage	Test Strand	Number of Items
Fiction ‡	Main ideas, details†	6
	Analyzes, interprets, synthesizes †	9
	Thinks critically*†	5
Nonfiction (Information or Task Oriented) ‡	Main ideas, details†	8
	Analyzes, interprets, synthesizes †	6
	Thinks critically*†	6
Total Number of Items		40

*Reading EALR 1: The student understands and uses different skills and strategies to read.

†Reading EALR 2: The student understands the meaning of what is read.

‡Reading EALR 3: The student reads different materials for a variety of purposes

Table 1-3: 2000 Grade 10, Number and Content of Writing Prompts

Task	Purposes[1]	Audiences[1]	Process[2]	Number of Prompts	Scores[3]
Extended Piece	Persuade	Editor	• prewrite • first draft • revise • edit • final draft	1	• Content, Organization & Style • Writing Mechanics
Extended Piece	Inform	Fellow Student	• prewrite • first draft • revise • edit • final draft	1	• Content, Organization & Style • Writing Mechanics
Total Number of Prompts				2	

[1] Writing EALR 1: The student writes clearly and effectively (concept & design, style [word choice, sentence fluency, voice], and conventions).

[2] Writing EALR 2: The student writes in a variety of forms for different audiences and purposes.

[3] Writing EALR 3: The student understands and uses the steps of a writing process*

Table 1-4: 2000 Grade 10, Number and Content of Mathematics Items

Process Strand	Concept Strand	Number of Items
Concepts & Procedures	Number Sense[1]	7
	Measurement[1]	6
	Geometric Sense[1]	5
	Probability and Statistics[1]	6
	Algebraic Sense[1]	5
Solves Problems[2]		4
Reasons Logically[3]		4
Communicates Understanding[4]		4
Making Connections[5]		5
Total No. of Items		46

[1] Mathematics EALR 1: The student understands and applies the concepts and procedures of mathematics.

[2] Mathematics EALR 2: The student solves problems using mathematics.

[3] Mathematics EALR 3: The student uses mathematical reasoning.

[4] Mathematics EALR 4: The student communicates knowledge and understanding in mathematical and everyday language.

[5] Mathematics EALR 5: The student makes mathematical connections.

ESTIMATED TESTING TIME PER SESSION—10th GRADE - SPRING 2000

The tests in the *Washington Assessment of Student Learning* are not timed. Students should have as much time as they need to work on the tests. Professional judgment should determine when a student is no longer productively engaged. When the majority of students have finished, the few still working may be moved to a new location to finish. Teachers' knowledge of students' work habits or special needs may suggest that some students who work very slowly should be tested separately or grouped with similar students for the entire assessment. For planning purposes, the estimated testing times required for most students are given in Table 1-5.

Table 1-5: Estimated Testing Times for Grade 10 WASL

Session	Subject	Approximate Time[1]
1	Listening	25 minutes
	Reading (Day One)	60 minutes
2	Reading (Day Two)	40 minutes
	Writing (Day One)	75 minutes
3	Writing (Day Two)	75 minutes
4	Mathematics (Day One) with tools	80 minutes
5	Mathematics (Day Two) without tools	80 minutes

[1] Above times are estimates for actual testing time. Additional time will be required to distribute and collect materials and cover the directions for test-taking. Testing sessions need not follow on consecutive days. Individual sessions should not be split but may be spaced with one or more days in between.

PART 2

TEST DEVELOPMENT AND CONTENT REPRESENTATION

The content of the *Washington Assessment of Student Learning* (WASL) state assessment is derived from the Washington state Essential Academic Learning Requirements (EALRs; see http://www.k12.wa.us/curriculuminstruct/EALRs.asp for the EALRs in all subject areas). These Essential Academic Learning Requirements define, for Washington schools, what students should know and be able to do by the end of grades 4, 7, and 10 in Reading, Writing, Communication, and Mathematics, and by the end of grades 5, 8, and 10 in Social Studies, Science, the Arts, Health and Fitness. The 2000 WASL tests measured EALRs for Reading, Writing, Mathematics, and Listening in grades 4, 7 and 10.

ITEM AND TEST SPECIFICATIONS

The first step in the test development process was to select the "Content Committees" that worked with staff of the Commission on Student Learning (CSL) and the Contractor (Riverside Publishing Company) to develop the actual items, which make up the assessments at each grade level. Each Content Committee was composed of 20 to 25 persons from around the state, most of whom were classroom teachers and curriculum specialists who had teaching experience at or near the grades and in the content areas that were to be assessed (i.e., Reading/Writing/Communication or Mathematics).

The second step in the development process was coming to a common agreement about the meaning and interpretation of the EALRs as well as which ones could be assessed on the state level test. During this step, it was very important that the Contractor, the Content Committees and the CSL staff were in agreement, in concrete ways, about what students were expected to know and be able to do and how these skills and knowledge would be assessed. In addition, the benchmark indicators were combined in various ways to create testing **targets** for which items would be written.

Next, Test and Item Specifications were prepared. Test Specifications define and describe such details as the kinds and number of items on the assessment, the blueprint or physical layout of the assessment, the amount of time to be devoted to each content area, and the scores to be generated once the test is administered. It was important that the goals of the assessment and the ways in which the results would be used be established at this stage so that the structure of the test would support the intended uses. In addition, the Test Specifications are the blueprint for developing equivalent test forms in subsequent years as well as creating new items to supplement the item pool. The final Test Specifications document the following topics:

- Purpose of the Assessment
- Strands
- Item Types
- General Considerations of Testing Time and Style
- Test Scoring
- Distribution of Test Items by Item Type

There are three types of items on the *Washington Assessment of Student Learning* (WASL) tests: multiple choice, short answer, and extended response. For each multiple-choice item, students select the one best answer from among three or four choices provided. Each multiple- choice item is worth one point. These items are machine scored.

The other two "open-ended" item types – short answer and extended response – require students to give their own responses in words, numbers, or pictures (including graphs or charts). Short-answer items are worth two points (scored 0, 1, or 2) and extended-response items are worth four points (scored 0, 1, 2, 3, or 4). For these items, student responses are assigned partial or full credit based on carefully defined scoring criteria. These items cannot be scored by machine and require hand-scoring by well-trained professional scorers (See Part 4).

In addition to the three item types, students are asked to complete two writing assignments (prompts). For grade 10, students write one informative piece and one persuasive piece. The writing prompts may require students to write a letter requesting information, describe an important event or situation, explain a procedure for completing a task or project. etc. Each written piece is worth six points and is hand-scored for content, organization, and style (1, 2, 3, or 4 points) and mechanics and spelling (0, 1, or 2 points).

Tables 2-1 through 2-3 are the test blueprints for item content and item types for the Reading, Listening, and Mathematics tests of the Grade 10 test. Based on the clarification of the EALRs and the Test Specifications, the next step was to develop Item Specifications. Item specifications provide sufficient detail, including sample items, to direct item writers in the development of appropriate test items for each assessment strand. Separate specifications were produced for the different types of items and for the different testing targets. The Test and Item Specifications documents were not only essential for WASL test construction but taken together they are powerful tools for teachers in developing their own assessments and for administrators in reviewing instructional programs. Test and Item Specifications are updated yearly, as needed. The most recent versions of these specifications can be obtained through the web site for the Washington State Office of the Superintendent of Public Instruction (OSPI): (See http://www.k12.wa.us/assessment/assessproginfo/default.asp for Test and Item Specifications in all subjects.).

Table 2-1: Grade 10 Reading Test: Item distribution by text type, strand, and item type

Text types/Strands	No. of Reading Selections	No. of Words Per Passage	No. of Multiple-Choice Items	No. of Short Answer Items	No. of Extended Response Items
Fiction‡	3	up to 1300	10-15	3-6	1
Comprehends important ideas and details†			3-5	1-2	0
Analyzes, interprets, synthesizes†			2-5	1-2	0-1
Thinks critically†*			2-5	1-3	0-1
Nonfiction‡	3-4	up to 1300	10-15	3-6	1
Comprehends important ideas and details†			3-5	1-2	0
Analyzes, interprets, synthesizes†			2-5	1-3	0-1
Thinks critically†*			2-5	1-3	0-1
Total	6-7	up to 4000	26-30	9-11	2

*Reading EALR 1: The student understands and uses different skills and strategies to read.

†Reading EALR 2: The student understands the meaning of what is read.

‡Reading EALR 3: The student reads different materials for a variety of purposes

Table 2-2: Grade 10 Listening Test: Item distribution by strand and item type

Strands	Number of Reading Selections	Number of Words Per Passage	Number of Multiple-Choice Items	Number of Short Answer Items
	2 editorials	up to 100	6-8	2
Listens and observes to gain and interpret information			3-5	0
Checks for understanding			2-3	1
Analyzes media messages			0-1	1
Total	2 editorials	up to 200	6-8	2

* Listening EALR 1: The student uses listening and observation skills to gain understanding.

Table 2-3: Grade 10 Mathematics Test: Item distribution by strand and item type

Strands	Multiple Choice	Short Answer	Extended Response
Number Sense[1]	3-7	1-2	0
Measurement Concepts[1]	3-7	1-2	0
Geometric Sense[1]	3-7	1-2	0
Probability and Statistics Procedures[1]	3-7	1-2	0
Algebraic Sense[1]	3-7	1-2	0
Solves Problems[2]	0-2	2-4	1-2
Reasons Logically[3]	0-2	1-4	0-1
Communicates Understanding[4]	0-2	1-4	0-1
Making Connections[5]	0-2	1-4	0-1
Maximum Number of Items	30	12	4
Maximum Number of Points	30	24	16

[1]Mathematics EALR 1: The student understands and applies the concepts and procedures of mathematics.

[2]Mathematics EALR 2: The student solves problems using mathematics.

[3]Mathematics EALR 3: The student uses mathematical reasoning.

[4]Mathematics EALR 4: The student communicates knowledge and understanding in mathematical and everyday language.

[5]Mathematics EALR 5: The student makes mathematical connections.

CONTENT REVIEWS

Once the Test and Item Specifications were completed and reviewed by the Content Committees, the Contractor's item writers prepared sample items and scoring criteria based on these specifications. Each Content Committee's task was then to review the items and scoring criteria to assure that the item writers had followed the specifications. As necessary items were revised to ensure that they measured Washington's Essential Academic Learning Requirements both accurately and comprehensively.

When the Content Committees were satisfied that the sample items and scoring criteria were appropriate, the item writers then produced literally hundreds of items to be pilot tested at the selected grade levels. Each test item was coded by content (EALR) area and item type (multiple choice, short answer, extended response) and presented to the Content Committees for final review just as they were to appear on the pilot test forms (including graphics, art work, and location on pages).

When the draft items were completed, the Content Committees reviewed each item, focusing on its fit to the Item Specifications, the EALRs, and the appropriateness of item content. For all short answer and extended response items, the proposed scoring guidelines (rubrics) were also reviewed. The Committees had three options with each item: approve the item (and scoring guidelines) as presented, recommend changes or actually edit the item (or scoring guidelines) to improve the item's "fit" to the EALRs and the Specifications, or eliminate the item from use in the assessment.

In addition to the Content Committees, a separate Fairness Review Committee reviewed each item to identify language or content that might be inappropriate or offensive to students, parents, or communities or items which might contain "stereotypic" or biased references to gender, ethnicity, or culture. As with the content reviews, The Fairness Review Committee reviewed each item and accepted, edited, or rejected it for use on the pilot assessment.

In order to be included on the pilot assessment, every item was reviewed by the Content Committees and the Fairness Review Committee. Approved items were to:

- be appropriate measures of the intended content;
- be appropriate in difficulty for the grade level of the examinees;
- have only one correct or best answer for each multiple-choice item;
- have appropriate and complete scoring guidelines for the open response items
- be free from content that might disadvantage some students for reasons unrelated to the concept or skill being tested

ITEM TRYOUTS

The approved items were then assembled into pilot test forms and administered to carefully-selected, representative samples of students across the state. All schools in the state of Washington were invited to participate in the pilot testing. Eighty five percent of fourth graders took part in the pilots. Test forms were randomly distributed with some effort to ensure that each test form was administered in districts with high populations of ethnic minority students. Each test form was administered to at least 1000 students.

SCORING AND ITEM ANALYSIS

Following the administration of the pilot assessment, student responses were evaluated by applying the scoring criteria approved by the Content Committees. A variety of statistical analyses were then employed to determine the effectiveness of the items and to check for item bias that may have been missed by the earlier reviews.

Two methods were used for item analysis. These were traditional or classical item analysis, which included the item means and item-test correlations for each item, and Rasch analysis, which included the item location and item fit. In addition, bias analysis was conducted using the Mantel-Haenszel bias statistic. Bias analysis investigates whether there is differential item performance for examinees of the same abilities who differ by virtue of gender or ethnicity.

Rasch Analysis

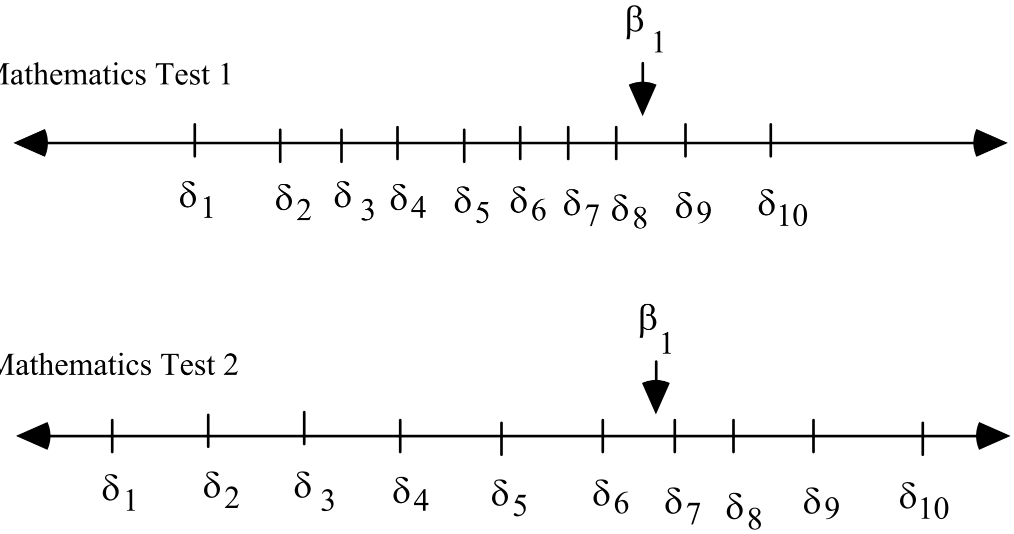
Rasch analysis is an Item Response Theory (IRT) analysis that places all items on a unique continuous scale for each content area. In addition, all examinees in the tryout pool are located on the same underlying scale. The Rasch analysis process separates item difficulty parameters from the abilities of the examinees in the sample that was tested. In this way, item difficulty parameters can be assumed to be the same for groups who are different from the original sample. The basic formula for the Rasch model is:

$$\mathrm{p}\left[x_{vi} = 1 \mid \beta_v, \delta_i\right] = \frac{\exp(\beta_v - \delta_i)}{1 + \exp(\beta_v - \delta_i)}$$

Where p = the probability of getting an item right given the ability of the examinee (βv) and the difficulty of the item (δi).

Working from this formula, item difficulties and examinee abilities can be estimated for a given test. The item difficulty location is the point on the ability scale where examinees have a 50/50 chance of getting an item correct. Figure 2-1 shows how examinee ability and item difficulty are placed on ability scales.

Figure 2-1: Location of examinee β_1 on two tests with item difficulties δ_1 through δ_{10}



Because the Rasch model can obtain an equal interval scale independent of item difficulty and person performance, the meaning of test scores can be interpreted in terms of scaled scores rather than number correct scores. For example, in Figure 2-1 (above), the examinee (β_1) got the first eight items correct on Mathematics Test 1 and the first six items right on Mathematics Test 2. The examinee is the same and her/his mathematics knowledge and skill remains the same; however, the ease or difficulty of the items result in different number-correct scores. The Rasch model will indicate the true distance of items from one another across the scale so that examinee test scores reflect the relative distance along the scale rather than the number of items answered correctly. The Rasch model separates item difficulty from examinee ability so that scores of examinees can be interpreted in terms of an underlying ability scale.

For items that have multiple points, a partial credit Rasch model is used to estimate the difficulty (threshold) of each *score* for an item. For example, items with 2 possible points can have two item thresholds: one for the point on the scale (location) at which examinees with abilities equal to that level on the scale have an equal chance of getting a score of 0 or 1, and one for the point on the scale at which examinees with abilities equal to that level on the scale have an equal chance of getting a score of 1 or 2. The formula for Master's partial credit model (which uses the Rasch dichotomous model as its base) is:

$$\pi_{xvi} = \frac{\exp \sum_{j=0}^{x} \left(\beta_v - \delta_{ij}\right)}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} \left(\beta_v - \delta_{ij}\right)} \qquad x = 0, 1, \ldots m_i$$

Where π equals the probability that an examinee with ability β_n will get score x on item i and δ_{ij} is the location of "step" j for item i (the point on the underlying scale where examinees have an equal probability of getting two adjacent scores [e.g., a score of 0 and a score of 1] on the item).

Once item scores are placed on a scale, items are assessed for "fit" to the Rasch model. The Rasch model assumes there was no guessing on multiple choice items and that, even though the items differ in terms of difficulty (or location on the scale) the items all function equally in discriminating between examinees below and above a given location on the scale. In order to be retained in the item pool, items must measure relevant knowledge and skill, represent desired locations on the ability scale, and fit the Rasch model.

Rasch analyses were conducted independently for each content area within the Washington Assessment of Student Learning (WASL). The fit of items depends upon whether the items in a scale were all measuring a similar body of knowledge and skill—in other words, whether the scale was unidimensional. Just as height, weight, and body temperature are different dimensions of the human body, so are Reading, Writing, Mathematics, and Listening different dimensions of achievement. Therefore, the items and scales for each test are examined independently.

In order to place all items across test forms on the same Rasch scale, a subset of items was repeated in adjacent forms. In other words, five items in Form 1 were repeated in Form 2; a different five items in Form 2 were repeated in Form 3; a different five items in Form 3 were repeated in Form 4; a different five items in Form 4 were repeated in Form 5; a different five items in Form 5 were repeated in Form 6; a different five items in Form 6 were repeated in Form 7; a different five items in Form 7 were repeated in Form 8 and a different five items from Form 8 were repeated in Form 1. In this way, Form 1 could be the anchor form and all items could be calibrated back to the item locations for the items in Form 1.

Traditional Item Analysis

For multiple-choice items, item means and item-test correlations constitute p-values and point-biserials respectively. These are the classical test theory equivalent of item difficulties and item discriminations. The p-value tells the percent of examinees who responded correctly to an item. Its value can range from 0 to 1.0. The point-biserial gives a measure of the relationship between performance on an item and performance on the test as a whole and can range from -1.0 to 1.0. For multiple-point (open-ended items), item means indicate the average earned score for examinees in the tryout sample. For 2-point items, item means can range from 0 to 2. For four-point items, item means can range from 0 to 4. Item-test correlations, for multiple point items, indicate the relationship between item performance and test performance. Item-test correlations can range from -1.0 to 1.0. Item-test correlations are computed using the test scores relevant to the item.

Unlike the Rasch item data, item means and item-test correlations are dependent on the sample of examinees that took the various tests. If the examinees were exceptionally well schooled in the concepts and skills tested, item means will be fairly high and the items will appear to be easy. If examinees are not well schooled in the concepts and skills tested, item means will be fairly low and items will appear to be difficult. If performance on an item does not relate well to performance on the test as a whole, item test correlations will be low or even negative. Hence both Rasch data and traditional item analysis data are used in item selection.

Bias Analysis

The Mantel Haenszel statistic is a chi-square (χ^2) statistic. Examinees are separated into relevant groups based on ethnicity or gender. Examinees in each group are ranked in terms of their total score on the relevant test. Examinees in the focal group (e.g., females) are compared with examinees in the reference group (e.g., males) in terms of their performance on individual items. Multiple 2x2 tables are created for each item (one for each total test score) indicating, for that score, the number of examinees in each group who got the item right and the number of examinees in each group who got the item wrong. Table 2-4 shows an example 2x2 table for performance on a hypothetical item for males and females with a total test score of 10 on a 40 point test. It appears that the item is more difficult for females than it is for males who had a total test score of 10.

Table 2-4: Responses to Item 3 for Males and Females with Total Test Score of 10

Item Number 3	Number Responding Correctly	Number Responding Incorrectly
Males (N = 100)	50	50
Females (N = 100)	30	70

Examinees with Total Test Score = 10

To complete the Mantel-Haenszel statistic, similar 2x2 tables are created for every test score. A χ^2 statistic is computed for each 2x2 table and the sum of all of the χ^2 statistics across all test scores gives the total bias statistic for a single item. When items have multiple points, a generalized Mantel-Haenszel statistic is computed using all points. Items that demonstrate a high $\Sigma\chi^2$ are flagged for potential bias. Generally, a certain percent of the items in any given pool of items will be flagged for item bias by chance alone. Careful review of items can help to identify whether some characteristic of an item may cause the bias (e.g., the content or language is unfamiliar to girls) or whether the bias data is probably a result of statistical error. For the WASL analyses, the alpha level (error level) was set at .01; therefore, about 1 percent of the items are expected to be flagged for bias by chance alone.

ITEM SELECTION

Statistical review of items involves examining item means, Rasch item difficulties (locations on the ability scale), and item-test correlations to determine whether items are functioning well. In addition, statistical review requires examining the "fit" of items to the Rasch model. Items that have extremely poor fit to the Rasch model must be revised or removed from the item pool prior to building a final test form. Items that function very poorly (are too easy, too difficult, or have low or negative item-test correlations) must also be revised or removed from the item pool. Finally, items that are flagged for bias against a focal group are examined closely to decide whether they will be removed from the pool. Generally, when item tryouts are conducted, sufficient numbers of items are developed so that revision and new tryouts are not needed. Faulty items can be deleted from the item pool.

After the statistical analyses were completed for the WASL, the Content and Fairness Review Committees reviewed these results and made the final determination about item quality and appropriateness based on the pilot test data. Items were reviewed again for fit to the EALRs; scoring rules were reviewed again for fit to the EALRs and to the demands of the items. In the Fairness Review Committees, bias data were reviewed to determine whether content or language may have resulted in large bias statistics. During these reviews, items were either accepted or rejected for the final pool of items.

Once these reviews were completed, the final pool of items was used to develop "operational" test forms. Operational test forms are those that are administered each year to monitor progress of schools and districts in helping students achieve the EALRs. Each operational form is developed by selecting items from the large pool of items tested in the 1998 item tryouts and approved by the Content and Fairness Review Committees. Four criteria are used in item selection for test forms:

1. Item quality
2. Content representation (See Test Specifications)
3. Representation of all gender and ethnic groups (See Test Specifications)
4. Item locations

Item quality is determined by the item means, item-test correlations, bias statistics, Rasch item locations, and fit statistics. Only the best items from the final pool are to be used in the operational test forms. Test specifications guide item selection to ensure that all relevant strands are represented in each test form as defined in the Test Specifications. Representation of all gender and ethnic groups is reviewed to ensure that Reading and Listening passages and stimulus materials used in the Mathematics and Writing tests give balanced representations of groups. Finally, because the WASL is intended to be a criterion-referenced test, and because performance standards are established for each test, items have been selected to represent a range of locations on the Reading, Mathematics, Writing, and Listening scales. After proficiency scores were established for each test in 2000 (See Part 5), item selection for subsequent years has ensured that item locations are similar to those in the initial operational test form in 2000.

Following the administration of the first operational Grade 10 assessment in Spring of 1999, the tests were scored for all participating students. A Standard-Setting Committee (see Part 5) was convened to establish the performance levels appropriate for reporting students' achievement of the EALRs. Based on the standards set by the Committee and approved by the Commission on Student Learning, results for the first Grade 10 operational assessment were reported in September, 1999. Table 2-5 gives the schedule of test development that was used for the Grade 10 WASL.

Table 2-5: Test Development Process for Grade 10

Action	Dates
Essential Academic Learning Requirements	March 1995
Test and Item Specifications	July-August 1997
Item Development	Sept.- Oct. 1997
Item Review (Content and Fairness)	November 1997
Pilot Testing	May 1998
Item Review (Content and Fairness)	Aug 1998
Item Bank	Sept 1998
Score Reports Designed	Sept 1998
Operational Tests Created	Oct - Dec 1998
Published Example Test Assessment Sampler	Feb 1999
First Operational Test Administered	April - May 1999
Standard Setting	June 1999

PART 3
EVIDENCE FOR THE VALIDITY OF INFERENCES FROM TEST SCORES

The most important issue in test development is the degree to which the achievement test actually elicits the conceptual understanding and skills that it is supposed to measure. In other words, when one claims that students must use logical reasoning skills to respond to an item, we need evidence that logical reasoning rather than memorization was actually used in the students' responses. Validity is an evaluative judgment about the degree to which the test ***scores*** can be interpreted to mean what test developers claim that they mean. Generally, there are about a half dozen different strategies for obtaining evidence for the validity of test scores (Messick, 1989):

1. We can look at the content of the test in relation to the content of the domain of reference;
2. we can probe the ways in which individuals respond to the items or tasks;
3. we can examine the relationships among responses to the tasks, items, or parts of the test, that is, the internal structure of test responses;
4. we can survey relationships of test scores with other measures and background variables, that is, the test's external structure;
5. we can investigate differences in these test processes and structures over time, across groups and settings, and in response to . . . interventions such as instructional . . . treatment and manipulation of content, task requirements, or motivational conditions;
6. finally, we can trace the social consequences of interpreting and using test scores in particular ways, scrutinizing not only the intended outcomes, but also the unintended side effects. (p. 16)

Validity, then, is a multidimensional construct that resides, not in tests, but in the relationships between any test score and its context (including the instructional practices and the examinee), the knowledge and skills it is to represent, the intended interpretations and uses, and the consequences of its interpretation and use. Messick stated that multiple sources of evidence are needed to investigate the validity of assessments. The following pages provide a description of the evidence available for the validity of scores on the Grade 10 *Washington Assessment of Student Learning* (WASL). This includes: correlations among scores and strands within the WASL; correlations between WASL tests and other achievement tests and measures of ability; and factor analysis studies examining evidence for the construct validity of WASL.

Part 2 of this technical report describes the process used in relation point 1 above: the judgment of content in relation to the subject area domains and selection of items that have adequate psychometric characteristics. While content representation and item quality are important aspects of tests, they do not ensure the validity of test scores. In order to examine the validity of test scores, it is important to determine whether examinees' performance within the set of items on the test is consistent (internal structure). This type of evidence is considered evidence for the construct validity of test scores. Studies to examine internal structure question whether the test scores elicit the constructs (knowledge and skills) the tests were intended to elicit.

Several studies have been conducted to gather evidence for the construct validity of the WASL Grade 10 Reading, Writing, Listening, and Mathematics Tests. The WASL Grade 10 1999 Technical Report provides information about studies that were conducted using 1999 Grade 4 WASL test data. In those reports, the results of studies conducted to gather evidence for the construct validity of the 1999 WASL Grade 10 Reading, Writing, Listening, and Mathematics Tests are presented. This involved examining the internal structure of the test scores, by looking at the intercorrelations among the items and strands assessed, and by conducting a factor analysis using the strand scores from the Grade 10 WASL scores. In this, the 2000 Grade 10 WASL Technical Report, the internal structure of the test has been reexamined including intercorrelations among WASL strand scores and factor analyses of WASL reading, writing, and mathematics strand scores. External evidence for validity has been examined through correlations among WASL test scores and subtest scores from the Iowa Test of Basic Skills (ITBS), and factor analyses of WASL strand score and subtest scores from the ITBS.

INTERNAL EVIDENCE FOR THE VALIDITY OF WASL SCORES

Correlations Among WASL Test Scores

The first analysis was that of correlations among WASL test scores. As can be seen in Table 3-1, responses to the different tests are moderately to strongly related. The strongest correlation is between scores on the WASL Reading Test and scores on the WASL Mathematics Test (.718). The next strongest are correlations between WASL Reading scores and WASL Writing scores (.693) and WASL Reading scores and WASL Listening scores (.652). Performance on the Listening Test was moderately correlated with performance on the Writing and Mathematics Tests. WASL Mathematics scores were moderately related to WASL writing scores.

Table 3-1: 2000 Grade 10 Correlations among WASL Test Scores

Tests	WASL Listening	WASL Reading	WASL Writing	WASL Mathematics
WASL Listening	1.00	.652	.525	.543
WASL Reading		1.00	.693	.718
WASL Writing			1.00	.625
WASL Mathematics				1.00

Intercorrelations among WASL Strand Scores

To more closely examine the relationships among performances on the WASL tests, the second analysis was of correlations among strand scores for Reading, Mathematics, and Writing, as well as the Listening Test scores. Table 3-2 gives the correlations among the strands within the 2000 WASL. As can be seen, scores for Reading strands (Main Ideas and Details of Fiction, Analysis, Interpretation, and Synthesis of Fiction, Critical Thinking about

Fiction, Main Ideas and Details in Nonfiction, Analysis, Interpretation, and Synthesis of Nonfiction, and Critical Thinking about Nonfiction) are moderately well correlated (.514 to .708) with the strongest correlations occurring between strands that measure analysis, interpretation, and synthesis of text and thinking critically about for all types of text. The Writing Content, Organization, & Style score is well correlated with the Writing Mechanics score (.602). Correlations among the Mathematics concepts scores (Number Sense, Measurement, Geometric Sense, Probability and Statistics, and Algebraic Sense) are moderately well correlated as would be expected given that these are diverse conceptual areas of Mathematics (.532 to .646). Prior research has shown that students perform differently on mathematical tasks that tap different areas of mathematics (Shavelson, Baxter, & Gao, 1993). The highest correlation is between Number Sense and Algebraic Sense (.646). Given that facility with numbers is required for both strands, this is to be expected.

Correlations among the Mathematical process scores (Solves Problems, Reasons Logically, Communicates Understanding, and Makes Connections) are also moderately strong (.590 to .659). The highest correlation is between scores for Solves Problems and scores for Communicates Understanding (.659). Since items for the Solves Problems strand are situated in contexts and require multiple content strands, this fairly high correlation is expected. The next highest correlation is between Solves Problems and Reasons Logically. It is likely that students must use many of their logical reasoning skills to solve problems.

Correlations between Mathematics content scores and Mathematics process scores are informative. Scores for Solves Problems, Reasons Logically, Communicates Understanding, and Makes Connections are moderately well correlated with scores for all content strands (.521 to .644). This suggests that content understandings are required for successful performance on all of the process strands. The highest correlations are between Communicates Understanding and Algebraic Sense (.644) and between Solves Problems and Algebraic Sense (.634). It may be appropriate to examine the kinds of communication required in the items measuring Communicates Understanding and Solves Problems to see if they require algebraic representations and the use of algebraic reasoning.

Correlations between Reading strand scores and Mathematics content strand scores are low to moderate (.334 to .500) with most between .40 and .50. The correlations between Reading strand scores and Mathematics process strand scores are also low to moderate (.365 to .575). The strongest relationships are between reading strand scores and scores for Solves Problems and Reasons Logically (.430 to .575). Although these are only moderate correlations, they may suggest that careful reading for details, interpretation and critical thinking are needed in items requiring reasoning and problem-solving. It is important to note that correlations between Writing strand scores and Mathematics strand scores are also low to moderate (.366 to .536). Writing strand scores also have only moderate correlations with all Reading strand scores, with most between .40 and .50. These correlations suggest that, for both the Reading Test and the Mathematics Test, skill in writing is only moderately related to performance.

Table 3-2: 2000 Grade 10 Correlations among Strands in the WASL

Strands	RL2	RL3	RI1	RI2	RI3	W1	W2	NS	ME	GS	PS	AS	SP	RL	CU	MC
Ideas & Details Fiction	.617	.550	.586	.574	.514	.424	.437	.395	.356	.402	.334	.381	.430	.456	.404	.365
Interpretation Fiction		.708	.644	.660	.606	.566	.543	.463	.411	.460	.380	.469	.514	.564	.492	.446
Critical Thinking about Fiction			.624	.648	.609	.592	.531	.465	.419	.457	.389	.478	.520	.575	.504	.453
Ideas & Details Nonfiction				.698	.636	.499	.515	.484	.447	.500	.418	.472	.526	.549	.498	.455
Interpretation Nonfiction					.621	.510	.523	.479	.432	.487	.406	.469	.521	.554	.496	.450
Critical Thinking about Nonfiction						.472	.457	.452	.419	.445	.388	.441	.488	.515	.468	.424
Content, Organization, Style							.602	.447	.406	.429	.380	.463	.483	.536	.480	.434
Writing Mechanics								.435	.378	.432	.366	.435	.453	.494	.444	.396
Number Sense									.609	.609	.575	.646	.626	.602	.635	.593
Measurement										.570	.532	.591	.604	.559	.606	.561
Geometric Sense											.539	.607	.607	.579	.600	.573
Prob. & Statistics												.555	.545	.521	.555	.521
Algebraic Sense													.634	.619	.644	.602
Solves Problems														.650	.659	.598
Reasons Logically															.647	.590
Communicates																.612

RL1-Main Ideas & Details of Fiction
RL2- Analysis, Interpretation, Synthesis of Fiction
RL3-Thinks Critically about Fiction
RI1-Main Ideas & Details of Nonfiction
RI2-Analysis, Interpretation, Synthesis of Nonfiction
RI3-Thinks Critically about Nonfiction
W1-Content, Organization, & Style
W2-Writing Mechanics

NS-Number Sense
ME-Measurement
GS-Geometric Sense
PS-Probability and Statistics
AS-Algebraic Sense

SP-Solves Problems
RL-Reasons Logically
CU-Communicates Understanding
MC-Makes Connections

Factor Analysis of WASL Listening Test Scores and Reading, Writing, and Mathematics Strand Scores

In order to follow up on these correlations, an exploratory factor analysis was conducted with the Listening Test scores, and the Writing, Mathematics and Reading strand scores. A principal components analysis was conducted using SPSS 8.0. The number of factors was determined using three criteria: eigen values greater than one, a scree plot, and the solution in which at least 63 percent of the variance was explained. Varimax (orthogonal) rotation was used. There were two plausible factor structures in the data. One (using the eigenvalues criterion) resulted in a two-factor solution that explained 58 percent of the total variance. Using a criterion of .60 for factor loadings (36% of the variance of a given variable), the two underlying factors were the language arts (Listening, Reading and Writing) and Mathematics. Table 3-3 gives the factor loadings (correlations between each of the variables and the underlying factors) from the rotated component matrix for the two-factor solution.

Table 3-3: 2000 Grade 10 Rotated Factor Loadings for Listening, Reading, Writing and Mathematics Strands for Two-Factor Solution

Variables	Language Arts Factor	Mathematics Factor
Listening	.716	.264
Main Ideas and Details of Fiction	.722	.208
Analysis, Interpretation, Synthesis of Fiction	.805	.282
Critical Thinking about Fiction	.773	.310
Main Ideas and Details of Nonfiction	.755	.332
Analysis, Interpretation, Synthesis of Nonfiction	.769	.316
Critical Thinking about Nonfiction	.711	.304
Content, Organization, and Style in Writing	.621	.362
Writing Mechanics	.626	.323
Number Sense	.302	.766
Measurement	.241	.754
Geometric Sense	.327	.717
Probability and Statistics	.217	.719
Algebraic Sense	.299	.768
Solves Problems	.386	.725
Reasons Logically	.481	.651
Communicates Understanding	.341	.754
Makes Connections	.283	.734

The second analysis resulted in a three-factor solution that explained 63 percent of the total variance. Using a criterion of .60 for factor loadings (36% of the variance of a given variable), the underlying factors were Reading and Listening, Mathematics, and Writing.

Table 3-4 gives the factor loadings (correlations between each of the variables and the underlying factors) from the rotated component matrix for the three-factor solution. While Reading, Writing, and Mathematics may be moderately correlated, Listening/Reading, Writing, and Mathematics strands represent separate dimensions of performance on the 2000 Grade 10 WASL as a whole. The fact that the Listening test scores load on the same factor as all Reading strand scores (albeit the lowest loading for that factor - .669) probably reflects the general comprehension required for both tests. The results presented here are consistent with results presented in the 1999 WASL Grade 10 Technical Report suggesting that the underlying factor structure for WASL is stable from year to year.

Table 3-4: 2000 Grade 10 Rotated Factor Loadings for Listening, Reading, Writing and Mathematics Strands for Three-Factor Solution

Variables	Mathematics Factor	Comprehension Factor	Writing Factor
Listening	.262	.683	.237
Main Ideas and Details of Fiction	.221	.756	.009
Analysis, Interpretation, Synthesis of Fiction	.271	.729	.350
Critical Thinking about Fiction	.292	.669	.405
Main Ideas and Details of Nonfiction	.339	.756	.173
Analysis, Interpretation, Synthesis of Nonfiction	.318	.750	.220
Critical Thinking about Nonfiction	.312	.722	.143
Content, Organization, Style in Writing	.300	.334	.770
Writing Mechanics	.264	.348	.749
Number Sense	.761	.264	.174
Measurement	.754	.230	.102
Geometric Sense	.715	.306	.141
Probability and Statistics	.719	.205	.010
Algebraic Sense	.757	.236	.226
Solves Problems	.719	.340	.209
Reasons Logically	.635	.391	.320
Communicates Understanding	.745	.283	.226
Makes Connections	.727	.240	.179

PERFORMANCE ACROSS GROUPS

Part 8 of this technical report presents data regarding performance of examinees across different categorical programs (i.e., Title I Reading, Title I Mathematics, LAP Reading, LAP Mathematics, S504, Special Education, Highly Capable Students, Bilingual/ESL, Title I Migrant). These data can be examined to determine whether patterns of performance are what would be expected on the test based on examinees' special needs. For example, students who have been identified as "highly capable" outperform all other groups on all tests. In addition, students who are in Title I Migrant and Bilingual/ESL programs have difficulty with reading and writing performance. Gender groups are also compared in Part 8. Whereas boys and girls perform equally well in Mathematics and Reading, girls outperform boys in Listening and Writing. These data, and other patterns in Tables 8-3 through 8-14, suggest that scores on the WASL tests are consistent with other measures of achievement in these subject areas.

SUMMARY

The results of these analyses provide evidence to support the validity of 2000 WASL scores. While achievement in one subject area is generally related to achievement in other subject areas, once WASL subscores are examined, it is evident that Listening and Reading, Mathematics, and Writing are different underlying dimensions of performance on the WASL tests.

Reference

Shavelson, R. J., Baxter, G. P., Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*, 215-232.

PART 4

SCORING THE WASL OPEN-ENDED ITEMS

During item development, scoring rubrics for each open-ended item on the *Washington Assessment of Student Learning* (WASL) were written (See http://www.k12.wa.us/assessment/assessproginfo/default.asp for Item Specifications). Item Specifications provide the general scoring rubrics that served as the guides for the item specific scoring rubrics for Reading, Mathematics, and Listening items. During item reviews, the scoring rubrics were reviewed along with item directions. A central aspect of the validity and reliability of test scores is the degree to which scoring rubrics are related to the appropriate learning targets (Essential Academic Learning Requirements) and whether they are applied faithfully during scoring sessions. The following procedures were used to score the WASL items. This information applies to all content areas that include open-ended items calling for student-constructed responses. These procedures were used for the full pool of items that were pilot tested, as well as for the 2000 operational tests.

QUALIFICATIONS OF READERS

Highly-qualified, experienced readers (scorers) were essential to achieving and maintaining consistency and reliability when scoring student-constructed (open-ended) responses. Readers selected for the Washington Assessment of Student Learning were required to have the following qualifications:

- A minimum of a bachelor's degree in an appropriate academic discipline (such as English, English Education, Math, Math Education, or related fields);
- Demonstrable ability in performance assessment scoring;
- Teaching experience, especially at the elementary or secondary level, was preferred.

Team and table leaders, responsible for supervising small groups of readers, were selected on the basis of demonstrated expertise in all facets of the scoring process, including strong organizational abilities, leadership, and interpersonal communication skills.

RANGE-FINDING AND ANCHOR PAPERS

The thoughtful selection of papers for range-finding and the subsequent compilation of anchor papers and other training materials were the essential first steps to ensure that scoring was conducted consistently, reliably, and equitably.

The range-finding process involved performance assessment specialists and curriculum specialists working with scoring team and table leaders as well as teachers from Washington state. All became thoroughly familiar with and reached consensus on the scoring rubrics approved by the Content Committees for each open-ended item. These range-finding teams began work with random selections of student responses for each item. They reviewed these responses, selected an appropriate range of responses, and placed them into packets which were numbered for easy reference. The packets of responses were read independently by members of a team of the most experienced readers. Following these independent readings and tentative ratings of the papers, the total range finding group worked together to discuss both the common and divergent scores. From this work, they assembled tentative sets of example responses for each prompt.

Then the primary task of the range-finding committee was to identify anchor papers – exemplars that clearly and unambiguously represented the solid center of a score point as described in the scoring criteria. Those exemplary anchor papers formed the basis not only of reader training, but of subsequent range-finding discussions.

Discussion was ongoing with the goal of identifying a sufficient pool of additional student responses for which consensus scores could be achieved and which illustrated the full range of student performance in response to the prompt or item. This pool of responses included borderline responses—ones that appeared to be between, rather than clearly within, a score level and which therefore represented a decision-making problem that readers (with training) would need to resolve.

TRAINING MATERIALS

Following the range-finding sessions, the performance assessment specialists and team leaders finalized the anchor sets and other training materials as identified in the range-finding meetings. The final anchor papers were chosen for their clarity in exemplifying the criteria defined in the scoring rubrics.

The anchor set for each 4-point item consisted of a minimum of thirteen papers, three examples of each of the four score levels and one example of a non-scorable paper. The anchor set for each 2-point item consisted of a minimum of seven papers, three examples of each score point and one example of a non-scorable paper. Score point exemplars consisted of one low, one solid mid-range, and one high example at each score point.

Additional training and qualifying sets of responses were selected to be used in reader training. One training set consisted of responses that were clear-cut examples of each score point; the second set consisted of responses closer to the borderline between two score points. The training sets gave readers an introduction to the variety of responses they would encounter while scoring, as well as allowing them to develop their decision-making capability for scoring responses that did not fall clearly into one of the scoring levels. Calibration/validity papers to be circulated during scoring were also identified at this time, as were reader-qualifying sets.

RATER CONSISTENCY (RELIABILITY)

Reader training for each prompt was led by a performance assessment specialists and team leaders. The primary purpose of the training was to help the readers understand the decisions made by the range-finding committee. Also, training helped readers internalize the scoring rubrics, so that they might effectively and consistently apply them.

Reader training sessions included an introduction to the assessment itself. In addition, readers were informed of the parameters or context within which the students' performance was elicited. This gave readers a better understanding of what types of responses could be expected, given such parameters as grade level, instruction, or time limitations. Readers next received a description of the scoring criteria that applied to the responses for each item.

The scoring criteria were always presented in conjunction with the anchor papers. After presentation and discussion of the anchor papers, each reader was given a training set consisting of ten papers. The readers scored the papers independently. When all readers had scored the training set, their preliminary scores were collected for reference.

Group discussion of the scores assigned was the next step, allowing the readers to raise questions about the application of the scoring rubric and giving them a context for those questions. The purpose of the discussion among the readers in training was to establish a consensus to ensure consistency of scores between readers. Even after readers had qualified to be scorers, training continued throughout the scoring of all responses to maintain high inter- and intra-reader reliability. Therefore, training was a continuous process and readers were consistently given feedback as they scored.

Frequent reliability checks were used to closely monitor the consistency of each reader's performance over time. The primary method of monitoring a reader's performance was by a process called "back-reading". In back-reading, each table leader reread and checked scores on an average of five to ten percent of each reader's work each day, with a higher percentage early in the scoring. If a reader was consistently assigning scores other than those the table leader would assign, the team leader and performance assessment specialist, together, retrained that reader, using the original anchor papers and training materials. This continuous, on-the-spot checking provided an effective guard against reader "drift," (beginning to score higher or lower than the anchor paper scores). Readers were replaced if they were unable to score consistently with the rubric and the anchor papers after significant training.

Tables 4-1 through 4-4 give the rater agreement information for the open-ended items in the 2000 Grade 10 WASL. Two types of rater agreement were calculated from 10 percent of the examinees randomly selected from the students' response booklets: score agreement for individual items and score agreement across the total score for the open-ended item set for each content area. For total score agreement on the open-ended items, the correlations were quite high (.95 to .99) within each content area with virtually no difference between the means of the total scores summed across open-ended items. For item-by-item interjudge agreement in Reading and Listening, the range of exact agreement was 80 to 97 percent and the range of exact and adjacent agreement was 97 to approximately 100 percent. For interjudge agreement in Writing, the range of exact agreement was 82 to 83 percent; exact and adjacent agreement was approximately 100 percent. For item-by-item interjudge agreement in Mathematics, the range of exact agreement was 90 to 99 percent and the range of exact and adjacent agreement was 99 to approximately 100 percent.

Table 4-1: 2000 Grade 10 Correlations between and Means of Total Scores of First and Second Readings for Open-Ended Items by Test.

Test	Correlation	Mean First Reading	Mean Second Reading
Listening & Reading	.99	18.22	18.06
Writing	.95	6.38	6.38
Mathematics	.99	15.41	15.39

Table 4-2: 2000 Grade 10 Frequencies of Exact Score Matches, Adjacent Scores, and Discrepant Scores for Listening and Reading Items.

Item	Points Possible	Exact Score Match	Adjacent Scores	Discrepant by Two Points	Discrepant by Three Points	Discrepant by Four Points	Percent Exact Agreement
3*	2	6886	750	27			90%
8*	2	6561	1088	14			86%
4	2	6905	748	10			90%
9	2	7450	170	43			97%
11	2	6839	797	27			89%
16	4	6139	1417	101	5	1	80%
17	2	6974	677	12			91%
20	2	6858	800	5			89%
24	2	7192	457	14			94%
26	4	6808	602	227	24	2	89%
28	2	7152	508	3			93%
31	2	6767	873	23			88%
34	2	6777	845	41			88%
40	2	7394	263	6			96%

* Listening items

Table 4-3: 2000 Grade 10 Frequencies of Exact Score Matches, Adjacent Scores, and Discrepant Scores for Writing Scores.

Score	Points Possible	Exact Score Match	Adjacent Scores	Discrepant by Two Points	Discrepant by Three Points	Percent Exact Agreement
1	4	3865	828	6		82%
2	2	3871	820	8		82%
3	4	3864	828	7		82%
4	2	3899	793	7		83%

Table 4-4: 2000 Grade 10 Frequencies of Exact Score Matches, Adjacent Scores, and Discrepant Scores for Mathematics Items.

Item	Points Possible	Exact Score Match	Adjacent Scores	Discrepant by Two Points	Discrepant by Three Points	Discrepant by Four Points	Percent Exact Agreement
3	2	13940	304	8	0	0	98%
6	4	13194	997	49	10	2	93%
9	2	13578	641	33	0	0	95%
11	2	13637	599	16	0	0	96%
15	2	13420	807	25	0	0	94%
16	4	13024	1106	108	12	2	91%
18	2	13410	829	13	0	0	94%
21	2	13542	702	8	0	0	95%
26	2	13781	462	9	0	0	97%
30	4	13108	1069	67	8	0	92%
33	2	13862	372	18	0	0	97%
36	2	13549	687	16	0	0	95%
38	4	13134	1078	37	3	0	92%
40	2	13472	765	15	0	0	95%
43	2	14171	73	8	0	0	99%
46	2	12764	1445	43	0	0	90%

ADDITIONAL CONSIDERATIONS FOR SCORING WRITING

Although the *training* for scoring writing is the same as described above, various approaches can be used in evaluation Writing. For the WASL, a "focused holistic" approach was selected. Focused holistic scoring, or general impression scoring, assesses relative writing fluency and measures the degree to which a writer has connected to the reader of a paper. When a paper is scored holistically, a reader considers the overall effectiveness of the piece of writing and assigns a score that reflects the reader's impression of the paper's overall quality. In a focused holistic approach, the reader also takes into account all of the elements that make up a successful piece of writing, for example content, organization, style, and mechanics. In the WASL Writing test, Content, Organization, and Style are scored together on a 4-point scale and Writing Mechanics are scored on a 2-point scale. These two scores are combined to provide a maximum of 6 points on any one piece of writing.

PART 5

STANDARD SETTING PROCEDURES

Standard setting for the Grade 10 *Washington Assessment of Student Learning* (WASL) was conducted in the summer of 1999. Because all of the items in the WASL item pool are on the same underlying Rasch scale (see Part 2), these standards can be held consistent across different test forms, making it possible to monitor student achievement over time with a fixed performance standard in each content area.

Standard setting committees were composed of teachers, curriculum specialists in the relevant subject area, school administrators, parents, and community members (Table 5-1). All standard setting committee members had direct experience with tenth graders or with the curriculum materials relevant for tenth graders.

Table 5-1: Number of Grade 10 Standard Setting Judges in each Professional Role.

Professional Role	**Number of Judges**
Language Arts Teachers*	14
Mathematics Teachers†	15
Specialist Teachers§	4
School Administrator§	6
Community Representative§	8
Total	47

* Reading, Writing, Listening Standard Setting Committee only

† Mathematics Standard Setting Committee only

§ Distributed across Mathematics and Reading, Listening, and Writing Standard Setting Committees

Setting standards for student performance on the WASL was essentially a systematic, judgmental process aimed at establishing a consensus, among knowledgeable people, about what tenth grade students should know and be able to do. Washington's Essential Academic Learning Requirements (EALRs) have described the expected content in Reading, Writing, Communications, and Mathematics for Washington's public schools (See http://www.k12.wa.us/curriculuminstruct/EALRs.asp). The new assessments have defined, in performance terms, some of the important knowledge, skills, and abilities tenth grade students should demonstrate in relation to the EALRs. The purpose of the standard-setting process was to establish the level of performance expected of tenth grade students who are judged as meeting the standards in Listening, Reading, Writing, and Mathematics. The emphasis for the judges, in the standard setting process, was on what students should know and be able to do near the end of Grade 10.

Performance standards on the Grade 10 assessment were determined by the standard setting procedure described below. This procedure is particularly well adapted to setting standards on assessments with mixed item types (that is, multiple-choice, short-answer, and extended response formats) as used on WASL. The procedure used in Washington state has been applied successfully in other large-scale assessment programs and was reviewed and approved by the National Technical Advisory Committee for the Commission on Student Learning—a committee composed of nationally recognized measurement professionals.

READING, LISTENING, AND MATHEMATICS

Implementation of the standard setting process required that the judges first take the operational test just as the students experienced it. The judges also reviewed scoring guides for the constructed-response (short-answer and open-ended) items and examples of student responses anchoring each item's score points.

Next, each standard setting judge received a complete set of the items ordered by difficulty from easiest to hardest, rather than in the order they appeared in the students' test booklets. Multiple-choice items appeared only once in the ordered booklet. Two-and four-point items appeared two or four times, according to the difficulty of achieving each score point. Data from the spring 1999 operational assessment was used to establish item difficulties. The first item in the judges' ordered booklet was the easiest item on the test, that is, the one the highest number of students answered correctly. The last item in the judges' ordered booklet was the hardest item on the test, that is, the one the fewest number of students answered correctly. Although the judges knew the items were ordered from easiest to most difficult, they did not know how students actually performed on the items—that is, how many students answered item 1 correctly, item 2 correctly, and so forth.

In small groups, the judges examined the items in the ordered booklet one at a time, starting with the first (easiest) item in the booklet, and moving to the second easiest item, and so on, until all items (and their scoring rubrics) were examined. As judges examined each item, they were asked to consider:

- What is each item measuring?
- What makes each item more difficult than the items that precede it?

Judges proceeded through the ordered item booklets and trained table leaders encouraged them to observe the increase in the complexity of the items and note the increase in knowledge, skills, and abilities required to answer the items.

At the conclusion of this first review of the ordered booklets, judges were asked to make an individual decision about where to place a "flag" at "meets standard". Each flag was placed in the ordered item booklet according to the individual judge's expectation of what students who are performing at standard should know and be able to do. For example, each judge placed his or her "meets standard" flag at a location in the booklet such that if a student is able to respond correctly to the items that precede the flag (with at least 2/3 likelihood of success), then the student has demonstrated sufficient knowledge, skills, and abilities to infer that the student is performing at the standard. For multiple-choice items this means the student who "meets standard" should be likely to know the correct response. For short answer- or extended response-items (with multiple score points), this means the student who "meets standard" should be likely to achieve at least that score point.

For the Reading and Mathematics tests, judges were asked to insert two additional flags: one at "exceeds standard" and one between "near standard" (partially proficient) and "low" (minimal). In this way, progress toward or beyond standards could also be identified. These additional flags were not set for the Listening test because there were not a sufficient number of points on each test to warrant such a fine distinction of performance levels.

Because not all judges set their flags in the same locations, the next step involved each judge sharing and discussing the locations at which his or her flag(s) were placed. When one judge placed a flag for "meets standard" farther along in the ordered booklet than another judge, it implied that the first judge expected students who meet the standard to demonstrate a higher level of achievement on the test. The difference in their individual expectations was reflected by the content and difficulty of the items between their flags.

For example, if Judge 1 placed a flag after item 30 and Judge 2 placed a flag after item 40, then these two judges disagreed on items 31-40. We know this because Judge 1, who placed a flag after item 30 was indicating that students who can correctly respond to the content in items 1-30 (with at least 2/3 likelihood) have demonstrated abilities sufficient to infer they have met the standard. Judge 2 (who placed the flag after item 40) did not agree, and was indicating that students have not demonstrated sufficient skills until they can handle more difficult content, that is, items 31-40.

Judges next discussed in small groups these differences in expectations as indicated by their different flag placements. Each group was provided with three lists indicating each judge's three flag locations for Reading and Mathematics. Beginning with the judges' placements of the "meets standard" flags, each judge was asked to note the location of every other judge's flag placement. Suppose the results in Table 5-2 occurred from the first round of standard setting.

Table 5-2: Example of Standard Setting Procedure

Judge Number	Meets Standard Flag Placed After:
1	item 30
2	item 34
3	item 29
4	item 33
5	item 36
6	item 39
7	item 33

Judges next would be asked to place a flag in their own ordered booklets after items 29, 30, 33, 34, 36, and 39. Now all judges could see the different expectations for student performance that "meet standard." In this example, judges would next discuss their differences, focusing on the items between 30 and 39 and discuss what these items ask of students' knowledge, skills, and abilities and whether students who meet the standard should be expected to respond correctly to these items. The discussion would consider the items one at a time beginning with item 30 and continuing up through item 39. When productive discussion of these items was completed, judges would then be asked to reevaluate their own initial flag locations in light of the small group discussion. Judges may decide to agree on a common flag placement during this round. That is, rather than requiring the calculation of the small group's average to determine the group's flag placement, the judges may agree to compromise and reach a consensus.

In the standard-setting for Reading and Mathematics, after judges had made their second round flag placements for "meets standard", the process was repeated for the other two cut-points—the below standard and the above standard locations.

Round 3 consisted of bringing the small groups back together as a large group to share and discuss each small group's flag placements. In the large group each judge placed a flag in his/her own ordered item booklet where each small group had made its flag placements. Large group discussion now focused on the items between the first and last flags for each performance level. Following the large group discussion, judges were asked to make a new (or reconfirm their former) flag placements.

Round 4 consisted of sharing with the large group the Round 3 small group results. Individual judges were then asked to make their final flag placements, which were then compiled to establish the final standard and other performance levels for each content area.

WRITING

Writing was handled in a slightly different manner than for Reading, Listening and Mathematics. There were two prompts (writing tasks). Each was scored for Content, Organization, and Style (1-4 points) and Mechanics (0-2 points). The scores from both prompts were combined (a possible range of 2-12 points) and the standard was set on the combined scores. To keep the standard-setting process for Writing as parallel with the other content areas as possible, the following standard-setting procedure was used:

1 Example responses were selected (both prompts together from the same student) that represented each of the possible combined score points 2-12 using a minimum of 3 students' responses for each possible score point.

2 These sets of combined student responses were ordered from lowest combined score (2) to highest combined score (12).

3 Judges were asked to proceed individually through all the example response sets (a minimum of 33) from lowest to highest and indicate the point at which the papers began to represent work "at the standard" and prior sets of papers represented work that was "less than the standard."

4 Next judges shared their individual judgments in their small groups and discussed the characteristics of the papers just above and just below their cut-points (flags).

5 The small group's placements were shared and discussed in the larger group.

6 Finally judges reconsidered their flags in light of the discussions and worked toward a consensus as to where the standard for Writing should be set.

SUMMARY

These processes ensured that the standards set for proficiency on the WASL tests would have careful scrutiny from a broad range of constituents of education. The judges had significant input from their peers and sufficient opportunities for discussion about their diverse opinions on standards.

PART 6
SCALE SCORES

All scaling for the Grade 10 *Washington Assessment of Student Learning* (WASL) was done using the same item data and calibrations used in the standard setting. Because the Mathematics and Reading Tests have four levels for student performance versus two levels for Listening and Writing, two different procedures were used to develop the scale scores. All four of the tests have a scale score of 400 representing the standard, but for Reading and Mathematics, the cut score for level two was set to equal 375 whereas in Listening and Writing an adjustment to the standard deviations was made to produce the scale scores. The following sections give details pertaining to the actual procedures used

DEVELOPMENT OF SCALE SCORES ON THE WASHINGTON ASSESSMENT OF STUDENT LEARNING

Scores on the WASL are reported as scale scores (See Tables 6-1 through 6-3 on Pages 8 through 10 of this chapter for 2000 Grade 10 number correct to scale scores conversions for each test.). As described in Part 2, the Rasch model and Master's (1982) extension of the Rasch model to multiple point items (the partial credit model) result in an equal interval scale (much like a ruler that is marked in inches or centimeters) for each test on which items and student scores can be reported. The partial credit model (PCM) allows for the inclusion of open-ended items where the maximum points possible are greater than one. Calibrating a test with Master's partial credit model produces estimated item parameters for an item's difficulty and the difficulty of its various score points (or steps). The possible scale score range for the WASL across the four test scales is 100 to 650 given all of the items in the item pool. This range is sufficient to describe levels of performance from the lowest possible earned scale score to the highest possible earned scale score *across all content areas tested and across different test forms*. The actual range of scale scores for each year and in each content area will differ. For example, the range of possible scale scores for the Grade 10 2000 Mathematics Test is 209 to 593 (See Table 6-3).

The Rasch model is an item response theory (IRT) model. IRT models can generate three parameters for items: item difficulties, item discriminations, and guess levels (the probability that low achieving examinees can guess correctly on multiple-choice items). The Rasch and PCM models also generate theta (θ) for each examinee. Because Rasch models treat all items as equally discriminating and assume that there is no guessing, there are no item discrimination and guessing parameters calculated. This means that, unlike more complicated scoring models, there is a one to one relationship between the number correct score on a test and the θ score on the test.

Once θ scores are generated, it is general practice to convert θ to a positive, whole number scale through a linear conversion procedure. The resulting numbers on the whole number scale are easy to use for computations when generating district, school, or building averages.

Because the scaled scores are on an equal interval scale, it is possible to compare score performance at different points on the scale. Much like a yard-stick, differences are constant at different measurement points. For example, a difference of 2 inches between 12 and 14 inches is the same differences as a difference of 2 inches between 30 and 32 inches. Two inches is two inches. Similarly, for equal interval achievement scales, a difference of 20 scaled score points between 360 and 380 means the same difference in achievement as a difference of 400 and 420, except that the difference is in degree of achievement rather than length.

The major limitation of scaled scores is that they are not well suited to making score interpretations beyond "how much more" and "how much less". Administrators, parents, and students ask, "What score is good enough? How do we compare with other schools like ours? Is a 40 point difference between our school and another school a meaningful difference?" For this reason, scale scores are usually interpreted by using performance standards or by converting them to percentile ranks.

Based on the content of the WASL, committees set performance standards for each subject area (Reading, Writing, Listening, and Mathematics) that would represent acceptable performance for a well-taught, hard working seventh grade student (see Part 4). In Reading and Mathematics, the standard setting committees also identified two "below standard" and one "above standard" performance levels[2]. Because the Listening and Writing Tests were relatively short, only two performance levels were established - "meets standard" and "does not meet standard."

The standard setting (described in Part 4) allowed the standard setting committees to identify the θ values associated with each cut-score (i.e., in Reading and Mathematics, the cut between "substantially below standard" and "approaches standard", between “approaches standard and” and "meets standard", and finally between "meets standard" and "exceeds standard"; in Writing and Listening, the cut between "does not meet standard" and "meets standard"). It was these θ values that formed the basics for the scaling procedure. In order to maintain the linear scale defined by the raw score to θ relationship, any two points on the θ

---

[2] The following are the general descriptions of the performance levels established for the Washington Assessment of Student Learning:

Level 4 -- Above Standard: This level represents superior performance, notably above that required for meeting the standard at grade 10.

Level 3 -- MEETS STANDARD: This level represents solid academic performance for grade 10. Students reaching this level have demonstrated proficiency over challenging content, including subject-matter knowledge, application of such knowledge to real world situations, and analytical skills appropriate for the content and grade level.

Level 2 -- Below Standard: This level denotes partial accomplishment of the knowledge and skills that are fundamental for meeting the standard at grade 10.

Level 1 -- Well Below Standard: This level denotes little or no demonstration of the prerequisite knowledge and skills that are fundamental for meeting the standard at grade 10.

In all content areas, the standard (Level 3) reflects what a well taught, hard working student should know and be able to do.

scale can be fixed to scale scores and the resulting transformation will remain linear. That is what was done here.

Reading and Mathematics

Following the standard setting process, a linear conversion was used to transform the θ (logistic ability) scores (from the PCM analyses) to a whole number scale. For all tests, the θ score identified as "meets standard" was converted to a WASL scale score of 400. For Reading and Mathematics, the θ score identified as "below standard-level 2" was converted to a Washington scale score of 375. The rest of the θ scores were converted to the whole number scale using the linear conversion equations for each test that produced these two scale score points. Since only two points can be set in a linear transformation, all other points must be derived from the conversion formula. Therefore, the "above standard" scale score for Reading was set at 414 and the "above standard" scale score for Mathematics was set at 427.

The general formula for a linear equation converting θ to a scaled score is:

$$\theta a + b = \text{scaled score} \qquad (6\text{-}1)$$

Where **a** is a distribution variable for the whole number scaled scores and **b** is a location on the whole number scale.

To obtain the linear formula necessary to translate from the θ scale to the whole number scale for Reading and Mathematics, the scaled score cut points for "meets standard" (400) and approaches standard (375) are plugged into the above formula and, through simultaneous solution of two equations, one can solve for **a** and **b.**

For math, the point on the θ scale where the standard setting committee decided that students had "met standard" was 0.286 and the point on the θ scale where the standard setting committee decided that students were "approaching standard" was -0.349. Therefore the initial linear equations were:

$$0.286\mathbf{a} + \mathbf{b} = 400 \qquad (6\text{-}2)$$
$$-0.349\mathbf{a} + \mathbf{b} = 375 \qquad (6\text{-}3)$$

Solving for a and b, the results are **a** = 39.37 and **b** = 388.74. These values were then used with the Mathematics θ scores to transform all θ scores to Mathematics scaled scores.

$$\text{Mathematics Scaled Score} = 39.37(\theta) + 388.74 \qquad (6\text{-}4)$$

For Reading, the point on the θ scale where the standard-setting committee decided that students had "met standard" was 0.793 and the point on the θ scale where the standard setting committee decided that students were "approaching standard" was –0.110. Therefore the initial linear equations were:

$$0.793\mathbf{a} + \mathbf{b} = 400 \tag{6-5}$$
$$-0.110\mathbf{a} + \mathbf{b} = 375 \tag{6-6}$$

Solving for a and b, the results are **a** = 27.69 and **b** = 378.05. These values were then used with the Reading θ scores to transform all θ scores to Reading scaled scores.

$$\text{Reading Scaled Score} = 27.69(\theta) + 378.04 \tag{6-7}$$

In Reading and Mathematics, students who earn scale scores below 375 are placed in the "below standard-level." Students who earn scale scores of 375 to 399 are placed in the "below standard, level 2" category in both Reading and Mathematics. Students who earn scale scores of 400 to 413 in Reading or 400 to 426 in Mathematics are in the "meets standard" category. Students who earn scale scores of 414 and higher in Reading or 427 and higher in Mathematics are in the "above standard" category.

Listening and Writing

In the standard setting for Listening and Writing only a single cut score was set representing the standard. Therefore the linear transformations θ for Listening and Writing required that one additional point be set. The decision was made to set the standard deviations of the θ scale of each test to a value so that the range of scale scores was within a 100 to 650 range obtained for the Reading and Mathematics Tests. Once the linear transformation formula was obtained, all θ for the Listening and Writing Tests were converted to whole number scaled scores. This means that scale scores of 400 or higher meet the standard in all content areas and scale scores of 399 or lower are below the standard.

CUT POINTS FOR CONTENT STRANDS

The cut points for the individual *content strands* in Reading and Mathematics were determined in the following manner. Using the θ value associated with "meets standard" and the item difficulties, it was possible to estimate the score of a proficient examinee on each of the items within the strand. Figure 6-1 gives a hypothetical distribution of item difficulties for the items in the Mathematics strands. As can be seen, the range of item difficulties differs for each strand. What may be less apparent is that the number of items below and above the theta value of .286 also differs. Students receiving raw scores for each of the strands equal to or higher than the estimated strand score for proficient examinees are reported as "similar to the performance expected of students who met the standard". Raw scores below this cut point are reported as "below the performance expected of students who met the standard". In Listening there are no scores reported at the strand level.

The Writing Test consists of only two writing prompts, so using the partial credit model is not appropriate. Instead all scaling was done on the raw score scale. In Writing the

Figure 6-1: Hypothetical Range of Item Difficulties (theta values) within Mathematics Strands

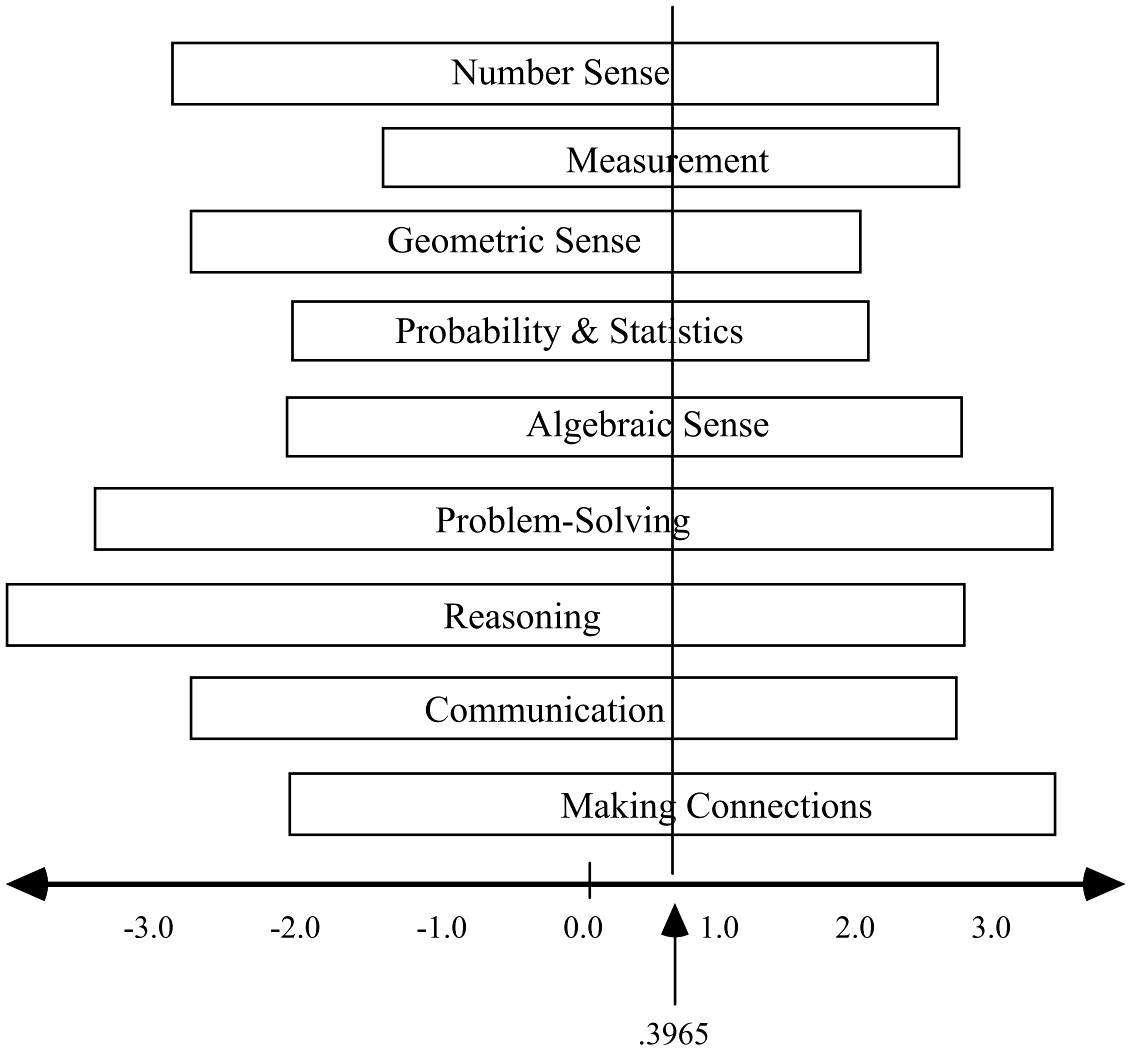
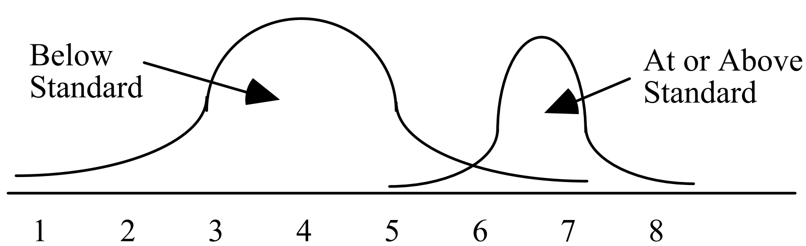


Figure 6-2: Score Distribution of Students Identified as Below Standard and Score Distribution of Students Identified to be at or Above Standard: Content, Organization, and Style



cut-score for the two strands were determined in the following manner. The data from the standard-setting was divided into two sets, one consisting of examinees meeting the standard, the other examinees not meeting the standard. The raw scores for Writing Content, Organization, and Style and for Writing Mechanics were obtained for the examinees in each group (those meeting the standard and those not meeting the standard). Frequency distributions were computed on each of the strands for each group. Cut-points were identified as those showing the smallest overlap between the distributions of the two groups (see Figure 6-2). This is often referred to as a "contrasting groups design". Discussions of the standard setting committees also contributed to the decision. In the end, a minimum combined score of six for the Writing Content, Organization, and Style strand and a combined score of three for the Writing Mechanics strand were determined to be the cut points and the item parameters.

EQUATING

The score scales established for the Grade 10 WASL in 1999 will stay in place for all subsequent years and test forms unless the scale is changed and new standards are set. Although new test forms are developed each year, Listening, Reading, and Mathematics will be equated using calibrations to items that were used in the base operational year (1999) – thus maintaining the same scale score system, i.e., 400 for meeting the standard. Although the raw score to scale score relationship will change for Listening, Reading, and Mathematics, the level of difficulty associated with meeting the standard in each tested content area will remain statistically equivalent over time.

The following is a summary of the procedures that are used for the equating of the Listening, Reading, and Mathematics Tests of the Grade 10 WASL. The same equating procedures and design was used for Grade 4 and Grade 7.

Equating Reading and Mathematics Tests

In the description that follows, the process was completed separately for Reading and Mathematics; however, because the Reading and Mathematics Tests are equated using the same design and procedure, the following description applies to both tests. In the first year of the operational assessment (1999), the multiple-choice, short-answer, and extended-response items were scaled using the Master's (1982) Partial Credit Model (PCM - see Pages 6-1 through 6-4 for a description of the model and the scaling process).

In order to equate the 1999 and 2000 test forms, anchor items were included in the Reading and Mathematics Tests of each form. These items were common from one form to the next. The first step in performing the equating procedure was to evaluate the stability of anchor items over time. All items for a test (e.g., Mathematics) in a given form were calibrated to a PCM scale. Item difficulty estimates for the anchor items within each test were obtained for the 1999 form (from item calibrations in the summer of 1999) and for the 2000 form. The mean of the item difficulties for the anchor items was computed separately for each test form. The difference between the means was computed to establish an "equating constant." The equating constant was added to the item difficulties of each of the anchor items from the 2000 scaling, thus resulting in equal means for the anchor items on the two test forms.

Next, the item difficulty for each anchor item from the 1999 scaling was subtracted from the adjusted item difficulty for the same anchor item from the 2000 scaling. Any item with an absolute difference greater than .3 was dropped from use as an anchor item (These items were not dropped from the test and from the generation of test scores, score reports, etc.; they were simply no longer to be used as anchor items.). If any items were dropped as anchor items, the computation of item difficulty means and equating constant, adjustment of item difficulties, and computation of differences in obtained and adjusted item difficulties was repeated. This process was repeated until the there was no loss of items.

Once a stable set of anchor items was obtained, the actual equating took place. This was done by analyzing the 2000 items for a test again, using the PCM, and fixing the item difficulties and step values for the valid anchor items to the values obtained on the 1999 test form. By fixing the item difficulties and step values of the anchor items, the resulting θ scale was the same in the 2000 test as it was for the 1999 test form. To derive the raw score to scale score relationship, the linear transformation equations for each test described on Pages 6-1 through 6-4 were used. This results in a consistent scale for each test across years.

Equating the Listening Test

Unlike the Reading and Mathematics Tests, the Listening Test is very short and consists of a single passage, read by the teacher, followed by six to eight items. This test design does not allow for the use of common items for equating. As a result, the contractor decided to use the anchor items from the Reading Test for equating. This made the equating of the Listening Test more complicated, involving more steps than needed for the Reading and Mathematics Tests. A key component of this analysis was to make sure that the integrity of the Listening scale was maintained despite the use of the Reading common items.

<u>Step 1</u>. To begin with, the item difficulties for the Listening items were obtained from the 1999 testing. Holding these item difficulties and step values fixed, the Listening Test items were rescaled including the Reading anchor items. This placed the Reading items on the Listening scale. This step was repeated for the 2000 test form, placing the Reading anchor items on the 2000 Listening scale.

<u>Step 2</u>. Using the Reading anchor item difficulties obtained in Step 1 for each form, it was possible to examine the stability of the common (anchor) items across forms. The same procedure outlined above for evaluating Reading and Mathematics anchor items was used to evaluate the Reading anchor items when projected onto the Listening scale.

<u>Step 3</u>. Once a set of stable anchor items was obtained for the 1999 to 2000 equating, the 2000 Listening Test items were analyzed using the PCM and holding the item difficulties and step values for the anchor items fixed to those found for the 1999 test form described in Step 1. This produced item difficulties and step values for the current Listening Test items that are on the same scale as the 1999 Listening scale.

<u>Step 4</u>. Using the item difficulties and step values obtained in Step 3, the raw score to θ scale values were obtained for the 2000 Listening Test.

Step 5. The final raw score to scale score relationship for the 2000 Listening Test was obtained by applying the same linear transformation used to obtain the raw score to scale score relationship for the 1999 form.

Equating the Writing Test

For Writing, writing prompts were selected for the 2000 WASL that were of similar difficulty, purpose and audience as those from the 1999 WASL (difficulty assessed based on tryout data). The same scoring criteria were used to ensure constancy in writing difficulty. There is no raw score to scale score table for the Writing Test. Writing scores are reported as raw scores.

NUMBER CORRECT SCORES TO SCALE SCORES

Each year WASL tests will have a different number correct score (raw score) to scale score relationship, although the underlying scale remains the same from year to year. This is possible because all items in the pool are on the same underlying Rasch scale. Table 6-1 gives the number correct score (NCS) to scale score (SS) relationship for the 2000 Grade 10 WASL Listening Test. Table 6-2 gives the NCS to SS relationship for the 2000 Grade 10 Reading Test. Table 6-3 gives the NCS to SS relationship for the 2000 Grade 10 Mathematics Test.

Table 6-1: 2000 Grade 10 Listening Number Correct Scores (NCS) to Scale Scores (SS)

NCS	Listening SS
0	241
1	276
2	317
3	345
4	369
5	390
6	410
7	431
8	456
9	494
10	529

Table 6-2: 2000 Grade 10 Reading Number Correct Scores (NCS) to Scale Scores (SS)

NCS	Reading SS	NCS	Reading SS
0	243	29	383
1	263	30	385
2	283	31	387
3	295	32	389
4	304	33	391
5	312	34	393
6	318	35	395
7	323	36	397
8	327	37	400
9	332	38	401
10	335	39	403
11	339	40	406
12	342	41	408
13	346	42	411
14	348	43	413
15	351	44	416
16	354	45	419
17	357	46	422
18	359	47	426
19	362	48	430
20	364	49	434
21	366	50	439
22	368	51	445
23	370	52	451
24	373	53	460
25	375	54	472
26	377	55	492
27	379	56	511
28	381		

Table 6-3: 2000 Grade 10 Mathematics Number Correct Scores (NCS) to Scale Scores (SS)

NCS	Mathematics SS
0	209
1	236
2	264
3	281
4	293
5	302
6	310
7	317
8	322
9	328
10	332
11	337
12	341
13	345
14	348
15	351
16	355
17	358
18	360
19	363
20	366
21	368
22	371
23	373
24	376
25	378
26	380
27	382
28	385
29	387
30	389
31	391
32	393
33	396
34	398
35	400

NCS	Mathematics SS
36	402
37	404
38	407
39	409
40	411
41	413
42	416
43	418
44	420
45	423
46	425
47	428
48	430
49	433
50	436
51	439
52	442
53	445
54	448
55	451
56	455
57	458
58	462
59	466
60	471
61	476
62	481
63	487
64	493
65	501
66	510
67	522
68	538
69	566
70	593

Reference

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrica*, (*47*), 149-174.

PART 7

RELIABILITY

The reliability of test scores is a measure of the degree to which the scores on the test are a "true" measure of the examinees' knowledge and skill relevant to the tested knowledge and skills. Simply put, the reliability is the proportion of observed score variance that is true score variance.

There are several ways to obtain estimates of score reliability: test-retest, alternate forms, internal consistency, and generalizability analysis are the most common. Test-retest estimates require administration of the same test at two different times. Typically the testing times for achievement tests are close together so that new learning does not impact scores. Alternate forms reliability estimates require administration of two parallel tests. These tests must be created in such a way that we have confidence that they measure the same domain of knowledge and skills using different items. Both test-retest and alternate forms estimates of the reliability of scores require significant testing time for examinees and are generally avoided when there is a concern that fatigue or loss of motivation might impact the resulting reliability coefficient.

The *Washington Assessment of Student Learning* (WASL) is a rigorous measure that requires significant concentration on the part of students for a sustained period of time. For this reason, it was determined that test-retest and alternate forms reliability methods were unlikely to yield accurate estimates of score reliability. Therefore, internal consistency measures were used to estimate score reliability for Reading, Listening, Writing, and Mathematics tests.

INTERNAL CONSISTENCY

Internal consistency reliability is an indication of how similarly students perform across items measuring the same knowledge and skills—in other words, how consistent each examinee performs across all of the items within a test. Internal consistency can be estimated using Cronbach's alpha coefficient. When a test is composed entirely of multiple-choice (dichotomously scored) items, a modification of Cronbach's alpha can be used (KR-20). However, when multiple-point items are included on a test, Cronbach's alpha coefficient provides the internal consistency estimate. Two of the demands for applying this method when estimating score reliability are: 1) the number of items should be sufficient to obtain stable estimates of students' achievement and 2) all test items should be homogeneous (similar in format and measuring very similar knowledge and skills).

WASL Reading and Mathematics tests have sufficient items to address the issue of test length; however, the Listening Test has fewer items/scores, hence this will have a tendency to depress the alpha coefficient. In addition, the Listening Test scores are generally high with a mean of 7.9 out of 10 possible points. This may also depress the alpha coefficient due to a restriction in the range of scores.

The WASL is also a complex measure that combines multiple-choice, short-answer, and extended response items. The Mathematics and Reading tests measure multiple strands that are all components of the domains of Mathematics and Reading respectively. Hence, examinee performance may differ markedly from one item to another due to prior knowledge, educational experiences, exposure to similar content, etc. Because of the heterogeneity of items in the Reading and Mathematics tests and the short test length for the Listening test, use of Cronbach's alpha coefficient for estimating score reliability for WASL will likely *under-estimate* of the actual reliability of scores. When items are heterogeneous, as they are in the WASL, it is generally believed that the true score reliability is higher than the estimate obtained through the alpha coefficient.

The WASL Writing Test is composed of two written essays. Although there are only four scores for the test (two for each of the essays), the items measure essentially the same ability twice. These items are very homogeneous; therefore, the alpha coefficient may be a reasonable estimate of the reliability of the scores.

The alpha coefficient is obtained through the following formula:

$$r_{xx'} = \left[\frac{N}{N-1}\right]\left(1 - \frac{\sum s_i^2}{s_x^2}\right)$$

Where:

$\sum s_i^2$ is the sum of all of the item variances

$\sum s_x^2$ is the observed score variance, and

N = the number of items on the test

Alpha coefficients for each of the 2000 Grade 10 WASL tests are given in Table 7-1. As can be seen, scores from the longer tests have higher reliability estimates. However, even with the very short Listening and Writing tests, these estimates provide good evidence for the overall reliability of 2000 Grade 10 WASL test scores.

Table 7-1: 2000 Grade 10 Reliability Estimates and Standard Error Of Measurement for Each WASL Test

Test	Alpha Coefficient	Scaled Score† or Raw Score Standard Error* of Measurement
Listening†	.56	35.6
Reading†	.87	10.9
Mathematics†	.91	12.0
Writing*	.79	1.10

STANDARD ERROR OF MEASUREMENT

One way to interpret the reliability of test scores is through the use of the Standard Error of Measurement (s_{EM}). The s_{EM} is an estimate of the standardized distribution of error around a given observed score. When one s_{EM} is added and subtracted from an observed score, we can be about 68 percent certain that the examinee's true score lies within the band. For example, the s_{EM} for the 2000 Grade 10 Reading Test is 10.89. If the examinee's scale score was 402, we could be about 68 percent certain that the examinee's true score was between 402 – 10.9 and 402 + 10.9, or between 391.1 and 412.9. If we add and subtract two s_{em}, we can be about 95 percent certain that the examinee's true score lies between 380.2 and 423.8. Finally, if we add and subtract three s_{em}, we can be about 99 percent certain that the examinee's true score lies between 369.3 and 434.7. In classical testing, we obtain the s_{em} through the following formula:

$$s_{em} = s_x \sqrt{1 - r_{xx'}}$$

Where:

s_x is the observed score standard deviation, and

$r_{xx'}$ is the reliability estimate (alpha)

Table 7-1 provides the 2000 Grade 10 standard error of measurement for the scaled scores of WASL Reading, Listening and Mathematics Tests based on the standard deviation of the scale scores and the alpha coefficient. Table 7-1 also gives the 2000 Grade 10 standard error of measurement for the raw scores of the WASL Writing Test based on the standard deviation of the raw scores and the alpha coefficient.

INTERJUDGE AGREEMENT

As was described in Part 4, inter-judge (inter-rater) agreement was another important source of evidence for the reliability of test scores. When two trained judges agree with the score given to a student's work, this gives support for the score on the short-answer or extended response item. Two methods are described in Part 4 for determining the degree to which judges gave equivalent score to the same student work: correlations between totals, when scores for open-ended items are summed, and percent agreement. For total score agreement on the open-ended items, the correlations were quite high (.95 to .99) within each content area with virtually no difference between the means of the total scores summed across open-ended items. For item-by-item interjudge agreement in Reading and Listening, the range of exact agreement was 80 to 97 percent and the range of exact and adjacent agreement was 97 to approximately 100 percent. For interjudge agreement in Writing, the range of exact agreement was 82 to 83 percent; exact and adjacent agreement was approximately 100 percent. For item-by-item interjudge agreement in Mathematics, the range of exact agreement was 90 to 99 percent and the range of exact and adjacent agreement was 99 to approximately 100 percent.

SUMMARY

In summary, the data from the interjudge agreement study indicates that the judges can consistently score performances using the scoring criteria developed for each item. Data from the alpha coefficients indicate that, except for the listening test, the test scores can be trusted to represent examinees' performance on the concepts and skills measured by the test. Standard errors of measurement, however, are large enough that caution should be used when evaluating and making decisions based on individual students' scores.

PART 8

DESCRIPTION OF PERFORMANCE FOR 2000 GRADE 10 STUDENTS

The data presented in this section of the report is descriptive of performance of fourth grade students throughout the state on the 2000 *Washington Assessment of Student Learning* (WASL). Included are means, standard deviations, and numbers tested for the all tested fourth graders and disaggregated by a variety of groups (Tables 8-1 through 8-14). Also presented are the percent of students in each gender, ethnic, and categorical program group who met or did not meet the standards for each content area (Tables 8-15 through 8-26). These data are useful for tracking, over time, the state's progress in helping students meet the Essential Academic Learning Requirements. *One possible limitation to the data is that the categorization of students is based on the way students are identified on their response books. If response books for given students did not indicate gender, ethnicity, and/or categorical program, the data for these students are not included in disaggregated data.* Finally, Tables 8-27 through 8-30 provide the mean performance on each item of the Grade 10 WASL tests, as well as the item-test correlations for each item.

SUMMARY STATISTICS

The means for each score were computed by summing the relevant scores for all students tested and dividing by the total number of students tested. The standard deviation was computed by obtaining the square root of the relevant variances using the following equation:

$$SD = \sqrt{\frac{\sum\left(X - \overline{X}\right)^2}{N}}$$

where:

X is the individual score

$\overline{X}$ is the mean of scores for all students tested in the state, and

N is the number of students tested in the state (those with valid scores)

Table 8-1 provides the state summary statistics for those Grade 10 students taking the WASL tests in 2000. The column headed "Points Possible" contains the maximum number of scale score points possible in each test for the 2000 form. The next two columns contain the mean scale score and standard deviation of the scale scores for all students tested in the state. Table 24 provides the state 2000 Grade 10 summary statistics for the WASL strands within tests. The column headed "Points Possible" indicates the maximum number of points possible in each strand for the 2000 form. The next two columns contain the mean number correct strand scores and standard deviation of the strand scores for all students tested in the state. The final column indicates the percent of students whose performance on the strand was similar to those who met the standard. Tables 25 through 28 provide the summary data for each ethnic and gender group tested in 2000 (as recorded on the response books). Table 29 through 32 provide the summary data for students in each of the following categorical programs: Learning Assistance Program (LAP) Reading, LAP Mathematics, Title 1 Reading, Title 1 Mathematics, Title 1 School, Bilingual/English as a Second Language (ESL), Highly

Capable Students, Section 504, Special Education, and Migrant Education (as recorded on the response books).

Table 8-1: 2000 Grade 10 Scale Score Means, Standard Deviations, and Maximum Scale Scores by Test

Test	Number Tested	Maximum Scale Score† or Raw Score*	Mean Scale Score or Raw Score	Standard Deviation
Listening†	68009	529	451.90	53.73
Reading†	67527	511	407.26	30.21
Writing*	64831	12	7.44	2.29
Mathematics†	68881	593	387.56	39.97

Table 8-2: 2000 Grade 10 Maximum Number Possible, Number Correct Score Means, Standard Deviations (SD) by Strand, and Percent of Students with Strength in Strand

Strand	Number with Valid Scores	Points Possible	Mean	SD	Percent with Strength in Strand
Main Ideas & Details of Fiction	67527	6	4.71	1.35	61.5
Analysis, Interpretation, Synthesis of Fiction	67527	12	7.96	2.65	59.3
Critical Thinking about Fiction	67527	10	5.93	2.65	55.5
Main Ideas & Details of Nonfiction	67527	9	7.06	1.99	48.9
Analysis, Interpretation, Synthesis of Nonfiction	67527	9	6.58	2.29	56.7
Critical Thinking about Nonfiction	67527	12	6.47	2.41	50.5
Writing Content, Organization Style	64831	8	4.73	1.33	26.5
Writing Mechanics	64831	4	2.71	1.23	54.2
Number Sense	68881	8	4.06	2.05	39.7
Measurement	68881	7	2.99	1.82	35.7
Geometric Sense	68881	7	3.94	2.01	40.5
Probability & Statistics	68881	7	3.10	1.71	38.2
Algebraic Sense	68881	7	2.70	1.95	47.1
Solves Problems	68881	9	3.66	2.30	35.4
Reasons Logically	68881	11	4.15	2.76	42.6
Communicates Understanding	68881	8	2.97	2.18	35.5
Makes Connections	68881	6	2.56	1.61	45.7

Table 8-3: 2000 Grade 10 Listening Test: Number Tested, Scale Score Means, and Standard Deviations (SD) by Gender

Gender	Number Tested	Mean	SD
Females	33306	459.57	50.86
Males	34531	444.60	55.34

Table 8-4: 2000 Grade 10 Listening Test: Number Tested, Scale Score Means, and Standard Deviations (SD) by Ethnic Group

Ethnic Group	Number Tested	Mean	SD
African American/Black	2701	426.79	59.99
Alaska Native/Native American	1368	433.09	55.66
Asian/Pacific Islander	5220	446.77	57.73
Latino/Hispanic	4604	421.09	61.49
White/Caucasian	50919	457.37	50.54
Multi-Ethnic	1061	446.10	50.79

Table 8-5: 2000 Grade 10 Reading Test: Number Tested, Scale Score Means, and Standard Deviations (SD) by Gender

Gender	Number Tested	Mean	SD
Females	33103	412.34	28.82
Males	34250	402.41	30.68

Table 8-6: 2000 Grade 10 Reading Test: Number Tested, Scale Score Means, and Standard Deviations (SD) by Ethnic Group

Ethnic Group	Number Tested	Mean	SD
African American/Black	2666	390.41	31.03
Alaska Native/Native American	1340	394.19	29.15
Asian/Pacific Islander	5192	406.90	31.01
Latino/Hispanic	4564	388.41	30.76
White/Caucasian	50595	410.42	29.02
Multi-Ethnic	1059	402.12	28.27

Table 8-7: 2000 Grade 10 Writing Test: Number Tested, Raw Score Means, and Standard Deviations (SD) by Gender

Gender	Number Tested	Mean	SD
Females	32174	7.98	2.10
Males	32504	6.90	2.35

Table 8-8: 2000 Grade 10 Writing Test: Number Tested, Raw Score Means, and Standard Deviations (SD) by Ethnic Group

Gender or Ethnic Group	Number Tested	Mean	SD
African American/Black	2464	6.46	2.30
Alaska Native/Native American	1241	6.49	2.24
Asian/Pacific Islander	4969	7.51	2.38
Latino/Hispanic	4183	6.06	2.30
White/Caucasian	48960	7.63	2.22
Multi-Ethnic	1005	6.90	2.30

Table 8-9: 2000 Grade 10 Mathematics Test: Number Tested, Scale Score Means, and Standard Deviations (SD) by Gender

Gender	Number Tested	Mean	SD
Females	33689	387.02	37.58
Males	35017	388.21	42.11

Table 8-10: 2000 Grade 10 Mathematics Test: Number Tested, Scale Score Means, and Standard Deviations (SD) by Ethnic Group

Ethnic Group	Number Tested	Mean	SD
African American/Black	2782	360.47	34.82
Alaska Native/Native American	1386	367.63	37.76
Asian/Pacific Islander	5283	394.10	40.22
Latino/Hispanic	4725	361.78	35.31
White/Caucasian	51500	391.58	39.06
Multi-Ethnic	1057	377.16	37.58

Table 8-11: 2000 Grade 10 Listening Test: Number Tested, Scale Score Means, and Standard Deviations (SD) by Categorical Program

Categorical Program	Number Tested	Mean	SD
LAP Reading	294	412.44	55.47
LAP Mathematics	313	417.32	60.76
Title 1 Reading	638	418.31	54.73
Title 1 Mathematics	566	420.92	54.55
Section 504	437	437.51	56.88
Special Education	4618	396.74	61.87
Title 1 Migrant Education	490	395.94	64.65
Bilingual/ESL	1709	381.09	60.86
Gifted/Highly Capable Students	1417	483.67	38.05

Table 8-12: 2000 Grade 10 Reading Test: Number Tested, Scale Score Means, and Standard Deviations (SD) by Categorical Program

Categorical Program	Number Tested	Mean	SD
LAP Reading	292	382.92	28.09
LAP Mathematics	311	381.89	28.90
Title 1 Reading	639	387.10	27.95
Title 1 Mathematics	563	388.68	27.74
Section 504	433	396.48	29.56
Special Education	4506	369.28	29.33
Title 1 Migrant Education	479	373.79	30.14
Bilingual/ESL	1703	368.38	29.09
Gifted/Highly Capable Students	1414	432.30	24.96

Table 8-13: 2000 Grade 10 Writing Test: Number Tested, Raw Score Means, and Standard Deviations (SD) by Categorical Program

Categorical Program	Number Tested	Mean	SD
LAP Reading	271	5.48	2.36
LAP Mathematics	286	5.50	2.30
Title 1 Reading	588	6.01	2.23
Title 1 Mathematics	523	6.14	2.26
Section 504	396	6.62	2.38
Special Education	4021	4.45	2.05
Title 1 Migrant Education	429	5.07	2.06
Bilingual/ESL	1424	4.54	2.01
Gifted/Highly Capable Students	1400	9.37	1.71

Table 8-14: 2000 Grade 10 Mathematics Test: Number Tested, Scale Score Means, and Standard Deviations (SD) by Categorical

Categorical Program	Number Tested	Mean	SD
LAP Reading	284	356.64	34.41
LAP Mathematics	321	352.01	33.65
Title 1 Reading	654	360.42	34.59
Title 1 Mathematics	590	360.32	34.88
Section 504	440	372.78	38.16
Special Education	4701	341.65	33.79
Title 1 Migrant Education	506	349.90	30.24
Bilingual/ESL	1757	355.97	35.38
Gifted/Highly Capable Students	1423	431.17	35.71

PERCENT MEETING STANDARD

Tables 8-15 through 8-22 provide the 2000 information regarding the number of students in each gender and ethnic group (as indicated on the response books) who met the standard in Listening, Reading, Writing, and Mathematics. Tables 23 through 30 provide the information regarding the number of students in each categorical program (as indicated on the response books) who met the standard in Listening, Reading, Writing, and Mathematics in 2000. The following are the general descriptions of the performance levels established for the Washington Assessment of Student Learning:

Level 4 Above Standard: This level represents superior performance, notably above that required for meeting the standard at grade 10.

Level 3 MEETS STANDARD*: This level represents solid academic performance for grade 10. Students reaching this level have demonstrated proficiency over challenging content, including subject-matter knowledge, application of such knowledge to real world situations, and analytical skills appropriate for the content and grade level.

Level 2 Below Standard: This level denotes partial accomplishment of the knowledge and skills that are fundamental for meeting the standard at grade 10.

Level 1 Well Below Standard: This level denotes little or no demonstration of the prerequisite knowledge and skills that are fundamental for meeting the standard at grade 10.

* *In all content areas, "Meets Standard" reflects what a well taught, hard working student should know and be able to do.*

For the Writing and Listening Tests, the tables show, for each group, the percent meeting standard and the percent not meeting standard. For the Reading and Mathematics tests, the tables show, for each group, the percent in each performance level. For Reading and Mathematics, students in Levels 1 and 2 did not meet the standard. Students in Levels 3 and 4 met or exceeded the standard.

Table 8-15: 2000 Grade 10 Listening Test: Percent Meeting Standards by Total Tested (N=68,009) and by Gender

Group	Percent Meeting Standard	Percent Not Meeting Standard	Percent Not Tested	Percent Exempt
All Students	84.0	16.0	7.3	1.8
Females	81.1	10.4	5.2	3.4
Males	72.4	16.5	6.5	4.5

Table 8-16: 2000 Grade 10 Listening Test: Percent Meeting Standards by Ethnic Group

Ethnic Group	Number of Students	Percent Meeting Standard	Percent Not Meeting Standard	Percent Not Tested	Percent Exempt
African American/Black	3314	57.7	23.8	10.9	7.6
Alaska Native/Native American	1663	61.8	20.5	11.4	6.3
Asian/Pacific Islander	5708	74.3	17.1	4.6	3.9
Latino/Hispanic	5446	55.7	28.8	9.2	6.3
White/Caucasian	55723	80.6	10.8	5.2	3.4
Multi-Racial	1124	79.0	15.4	4.7	0.9

Table 8-17: 2000 Grade 10 Reading Test: Percent Meeting Standards by Total Tested (N=67,527) and by Gender

	Meets Standard		Does Not Meet Standard			
Group	Percent Level 4	Percent Level 3	Percent Level 2	Percent Level 1	Percent Not Tested	Percent Exempt
All Students	47.9	14.5	18.9	10.9	7.8	1.8
Females	43.7	21.8	17.4	8.0	5.8	3.4
Males	31.1	21.6	20.9	14.5	7.3	4.5

Table 8-18: 2000 Grade 10 Reading Test: Percent Meeting Standards by Ethnic Group

		Meets Standard		Does Not Meet Standard			
Ethnic Group	Number of Students	Percent Level 4	Percent Level 3	Percent Level 2	Percent Level 1	Percent Not Tested	Percent Exempt
African American/Black	3314	17.0	18.2	23.7	21.5	12.1	7.5
Alaska Native/Native American	1663	19.3	19.0	23.8	18.5	13.2	6.3
Asian/Pacific Islander	5708	37.7	20.9	19.4	12.9	5.3	3.8
Latino/Hispanic	3314	17.0	18.2	23.7	21.5	12.1	7.5
White/Caucasian	55723	41.2	22.5	18.1	9.0	5.8	3.4

Multi-Racial	1124	30.4	25.4	24.8	13.5	4.9	0.9

Table 8-19: 2000 Grade 10 Writing Test: Percent Meeting Standards by Total (N=64,831) and by Gender

Group	Percent Meeting Standard	Percent Not Meeting Standard	Percent Not Tested	Percent Exempt
All Students	46.9	53.1	11.1	1.7
Females	40.1	48.2	8.3	3.4
Males	22.9	60.8	11.8	4.5

Table 8-20: 2000 Grade 10 Writing Test: Percent Meeting Standards by Ethnic Group

Ethnic Group	Number of Students	Percent Meeting Standard	Percent Not Meeting Standard	Percent Not Tested	Percent Exempt
African American/Black	3314	15.7	58.7	17.9	7.7
Alaska Native/Native American	1663	15.3	59.3	18.9	6.4
Asian/Pacific Islander	5708	34.1	52.9	9.1	3.9
Latino/Hispanic	5446	11.8	65.0	16.7	6.5
White/Caucasian	55723	34.4	53.4	8.8	3.4
Multi-Racial	1124	24.6	64.8	9.7	0.9

Table 8-21: 2000 Grade 10 Mathematics Test: Percent Meeting Standards by Total (N=68,881) and Gender

	Meets Standard		Does Not Meet Standard			
Group	Percent Level 4	Percent Level 3	Percent Level 2	Percent Level 1	Percent Not Tested	Percent Exempt
All Students	19.0	19.9	20.5	32.4	8.2	1.7
Females	13.0	20.8	24.7	34.0	4.2	3.3
Males	16.2	19.3	20.6	34.1	5.4	4.4

Table 8-22: 2000 Grade 10 Mathematics Test: Percent Meeting Standards by Ethnic Group

		Meets Standard		Does Not Meet Standard			
Group	Number of Students	Percent Level 4	Percent Level 3	Percent Level 2	Percent Level 1	Percent Not Tested	Percent Exempt
African American/Black	3314	2.8	8.0	16.8	56.3	8.5	7.6
Alaska Native/Native American	1663	5.1	11.1	18.7	48.5	10.3	6.3
Asian/Pacific Islander	5708	19.1	21.3	22.4	29.6	3.8	3.6
Latino/Hispanic	5446	3.7	8.1	16.9	58.1	7.1	6.2
White/Caucasian	55723	16.5	22.2	23.6	30.2	4.2	3.3
Multi-Racial	1124	9.0	14.2	24.7	46.1	4.8	1.2

Table 8-23: 2000 Grade 10 Listening Test: Percent Meeting Standards by Categorical Program

Categorical Program	Number of Students	Percent Meeting Standard	Percent Not Meeting Standard	Percent Not Tested	Percent Exempt
LAP Reading	338	55.9	31.1	8.0	5.0
LAP Mathematics	366	54.9	30.6	10.1	4.4
Title 1 Reading	726	57.9	30.0	7.9	4.3
Title 1 Mathematics	653	58.7	28.0	9.3	4.0
Section 504	488	68.4	21.1	5.7	4.7
Special Education	6039	38.3	38.2	11.3	12.2
Title 1 Migrant Education	561	42.8	44.6	8.0	4.6
Bilingual/ESL	2163	30.6	48.4	7.6	13.4
Gifted/Highly Capable Students	1446	96.0	2.0	1.7	0.3

Table 8-24: Grade 10 Reading Test: Percent Meeting Standards by Categorical Program

		Meets Standard		Does Not Meet Standard			
Categorical Program	Number of Students	Percent Level 4	Percent Level 3	Percent Level 2	Percent Level 1	Percent Not Tested	Percent Exempt
LAP Reading	338	10.9	13.6	31.1	30.8	8.6	5.0
LAP Mathematics	366	10.4	12.8	31.1	30.6	10.4	4.6
Title 1 Reading	726	14.2	17.9	28.4	27.5	7.9	4.1
Title 1 Mathematics	653	14.9	18.4	28.2	24.8	9.8	4.0
Section 504	488	21.7	21.3	27.3	18.4	7.0	4.3
Special Education	6039	3.9	8.8	20.7	41.2	13.2	12.2
Title 1 Migrant Education	561	6.6	11.1	27.1	40.6	10.0	4.6
Bilingual/ESL	2164	4.8	7.5	20.4	46.0	7.9	13.4
Gifted/Highly Capable	1466	76.4	15.5	4.4	1.5	1.9	0.3

Table 8-25: 2000 Grade 10 Writing Test: Percent Meeting Standards by Categorical Program

Categorical Program	Number of Students	Percent Meeting Standard	Percent Not Meeting Standard	Percent Not Tested	Percent Exempt
LAP Reading	338	9.8	70.4	14.8	5.0
LAP Mathematics	366	9.0	69.1	16.9	4.9
Title 1 Reading	726	11.0	70.0	14.5	4.5
Title 1 Mathematics	653	12.9	67.2	15.3	4.6
Section 504	488	20.1	61.1	13.9	4.9
Special Education	6039	2.7	63.9	21.1	12.3
Title 1 Migrant Education	561	3.9	72.5	18.2	5.3
Bilingual/ESL	2164	2.9	62.9	20.2	14.0
Gifted/Highly Capable Students	1446	71.1	25.7	2.8	0.3

Table 8-26: 2000 Grade 10 Mathematics Test: Percent Meeting Standards by Categorical Program

		Meets Standard		Does Not Meet Standard			
Categorical Program	Number of Students	Percent Level 4	Percent Level 3	Percent Level 2	Percent Level 1	Percent Not Tested	Percent Exempt
LAP Reading	338	3.3	5.6	12.4	62.7	11.2	4.7
LAP Mathematics	336	1.4	5.7	12.6	68.0	7.7	4.6
Title 1 Reading	726	2.8	9.0	16.3	62.1	6.2	3.7
Title 1 Mathematics	653	2.8	9.3	16.2	62.0	5.4	4.3
Section 504	488	8.4	11.5	19.9	50.4	5.3	4.5
Special Education	6039	1.0	2.9	7.6	66.3	9.7	12.4
Title 1 Migrant Education	561	1.1	3.9	12.5	72.7	5.2	4.6
Bilingual/ESL	2164	2.9	6.2	12.8	59.3	5.8	13.0
Gifted/Highly Capable	1446	56.0	25.4	11.5	5.4	1.3	0.3

MEAN ITEM PERFORMANCE AND ITEM-TEST CORRELATIONS

As discussed in Part 2, traditional item statistics were used, along with Rasch difficulties and fit statistics, to evaluate the quality of items. All items in the pool were evaluated together and items that met quality standards were retained in the item pool. Mean item performance for multiple choice items can range from 0 to 1. This is often called the p-value. Mean item performance for short-answer items can range from 0 to 2. Mean item performance for extended response items can range from 0 to 4. For the Writing test, mean scores represent the average scores for each of the scoring rules applied to the written piece. There are two written pieces in the Grade 10 WASL. Students can receive from 0 to 4 points for Content, Organization, and Style and from 0 to 2 points for Writing Mechanics for *each* of the written pieces. The higher the mean item performance, the easier the item. Item-test correlations can range from -1.0 to 1.0; positive correlations indicate that item performance is related to overall test performance. Rasch item difficulties can range from –4.0 to 4.0, with negative numbers representing easier items and positive numbers representing more difficult items. The data provided in Tables 8-27 through 8-30 indicate the number of points possible for the items or writing scores, the item or score means, the item score to test score correlations, and the Rasch item difficulties for each of the items in the Listening, Writing, Reading, and Mathematics tests respectively.

Table 8-27: 2000 Grade 10 Listening Test: Number of Points Possible Per Item, Mean Item Performance, Item-Test Correlation, and Rasch Item Difficulty for Each Item

Item Number in Test Booklet	Number Possible	Item Mean	Item-Test Correlation	Rasch Item Difficulty
1	1	0.95	0.34	-0.92
2	1	0.93	0.35	-0.71
3	2	1.34	0.43	1.39
4	1	0.74	0.06	0.90
5	1	0.59	0.13	1.59
6	1	0.77	0.32	0.70
7	1	0.82	0.27	0.40
8	2	1.04	0.24	1.80

Table 8-28: 2000 Grade 10 Writing Test: Number of Points Possible Per Score-Type, Mean Score, and Score-Total Test Correlation for Each Score

Prompt Number	Score Type	Score Points Possible	Score Mean	Score-Total Test Correlation
1	Content, Organization & Style	4	2.28	0.58
	Writing Mechanics	2	1.30	0.61
2	Content, Organization & Style	4	2.34	0.56
	Writing Mechanics	2	1.34	0.59

Table 8-29: 2000 Grade 10 Reading Test: Number of Points Possible Per Item, Mean Item Performance, Item-Test Correlation, and Rasch Item Difficulty for Each Item

Item Number in Test Booklet	Points Possible	Item Mean	Item-Test Correlation	Rasch Item Difficulty
1	1	0.82	0.47	-0.83
2	1	0.67	0.38	0.16
3	1	0.69	0.36	0.04
4	2	1.20	0.52	0.54
5	1	0.63	0.25	0.38
6	1	0.89	0.32	-1.45
7	1	0.71	0.31	-0.11
8	1	0.58	0.37	0.58
9	2	1.15	0.32	0.58
10	1	0.80	0.32	-0.61
11	2	1.16	0.53	0.66
12	1	0.81	0.47	-0.75
13	1	0.88	0.51	-1.33
14	1	0.88	0.47	-1.36
15	1	0.76	0.47	-0.38
16	4	2.01	0.65	1.00
17	2	1.23	0.55	0.38
18	1	0.76	0.36	-0.37
19	1	0.82	0.49	-0.83
20	2	1.16	0.59	0.53
21	1	0.81	0.49	-0.73
22	1	0.78	0.32	-0.52
23	1	0.72	0.49	-0.12
24	2	1.48	0.60	-0.02
25	1	0.89	0.43	-1.54
26	4	2.80	0.68	0.22
27	1	0.84	0.52	-1.02
28	2	1.09	0.57	0.74
29	1	0.70	0.29	-0.04
30	1	0.79	0.44	-0.60
31	2	1.33	0.61	0.29
32	1	0.67	0.32	0.18
33	1	0.47	0.28	1.14
34	2	1.09	0.62	0.85
35	1	0.55	0.27	0.73
36	1	0.65	0.44	0.20
37	1	0.85	0.47	-1.06
38	1	0.69	0.41	0.06
39	1	0.84	0.45	-0.77
40	2	1.51	0.53	-0.12

Table 8-30: 2000 Grade 10 Mathematics Test: Number of Points Possible Per Item, Mean Item Performance, Item-Test Correlation, and Rasch Item Difficulty for Each Item

Item Number in Test Booklet	Points Possible	Item Mean	Item-Test Correlation	Rasch Item Difficulty
1	1	0.56	0.39	-0.47
2	1	0.52	0.54	-0.13
3	2	1.44	0.44	-1.04
4	1	0.29	0.34	1.05
5	1	0.36	0.32	0.65
6	4	2.11	0.69	-0.19
7	1	0.34	0.28	0.76
8	1	0.48	0.36	0.06
9	2	0.74	0.52	0.60
10	1	0.41	0.36	0.39
11	2	0.96	0.49	0.05
12	1	0.66	0.43	-0.80
13	1	0.56	0.53	-0.23
14	1	0.17	0.20	1.79
15	2	0.88	0.56	0.20
16	4	0.73	0.51	1.45
17	1	0.50	0.39	0.01
18	2	0.86	0.57	0.00
19	1	0.69	0.51	-1.01
20	1	0.59	0.54	-0.47
21	2	0.65	0.63	0.79
22	1	0.48	0.35	0.07
23	1	0.58	0.33	-0.41
24	1	0.52	0.53	-0.16
25	1	0.51	0.42	0.01
26	2	0.91	0.52	-0.10
27	1	0.54	0.31	-0.25
28	1	0.32	0.30	0.88
29	1	0.26	0.21	1.19
30	4	2.02	0.62	0.04
31	1	0.48	0.34	0.27
32	1	0.40	0.31	0.46
33	2	0.86	0.51	0.20
34	1	0.54	0.57	-0.25
35	1	0.26	0.40	1.19
36	2	1.17	0.49	-0.48
37	1	0.19	0.44	1.63
38	4	1.19	0.71	0.79
39	1	0.48	0.37	0.04
40	2	0.74	0.32	0.73
41	1	0.29	0.26	1.03
42	1	0.49	0.33	0.01
43	2	0.30	0.48	1.47
44	1	0.31	0.48	0.93
45	1	0.42	0.41	0.33
46	2	0.90	0.49	0.17

APPENDIX A

WASHINGTON ASSESSMENT OF STUDENT LEARNING

National Technical Advisory Committee

Washington State Technical Advisory Committee

National Technical Advisory Committee Members

Patricia Almond, Education Specialist, Evaluation, Oregon Department of Education

Peter Behuniak, Director of Testing, Connecticut State Department of Education

Richard Duran, Professor, University of California – Santa Barbara

Robert Linn, Professor, University of Colorado and UCLA/CRESST

William Mehrens, Professor, Michigan State University

Joseph Ryan, Professor, Arizona State University

Kenneth Sirotnik, Professor, University of Washington

Catherine Taylor, Associate Professor, University of Washington

Martha Thurlow, Director of the Center for Education Outcomes, University of Minnesota

Washington State Assessment Advisory Team

Charisse Berner, Curriculum Director, Oak Harbor School District

Linda Elman, Director of Research and Evaluation, Central Kitsap School District

Robert Hamilton, Director of Research and Evaluation, Northshore School District

Bev Henderson, Director of Assessment and Staff Development, Kennewick School District

Peter Hendrickson, Director of Research and Evaluation, Evergreen School District

Joe Kinney, Director of Research and Evaluation, Spokane School District

Duncan MacQuarrie, Director of Assessment, Tacoma School District

Mike O'Connell, Director of Assessment and Research, Seattle Public Schools

Dan Penhallegon, Director of Research and Evaluation, Yakima School District

Nancy Skerritt, Assistant Superintendent for Curriculum and Assessment, Tahoma School District

Rick Williams, Director of Research and Evaluation, Everett School District

Phil Dommes, Director of Assessment and Evaluation, North Thurston Public Schools